Table of Contents

SUPPLEMENTARY NOTES

Supplementary Note 1 Ecological and genetic background of the sequenced Atlantic cod specimen

The North East Arctic cod (NEAC) and the Norwegian coastal cod (NCC) represent the two major populations of Atlantic cod in the Norwegian waters. The stationary NCC is thought to be further structured into several local stocks along the coast of Norway (Knutsen et al. 2007), while the NEAC (or "skrei") migrates from the Barents Sea to the main spawning ground of Lofoten. Although interbreeding may occur in this region, the two populations are differentiated based on certain genetic markers such as pantophysin (PanI, Jakobsdóttir et al. 2011). The North East Arctic cod population is of most significance for Norwegian fisheries and is considered the largest population of cod in the North East Atlantic ocean. Currently, the spawning stock is estimated at 1.35 million tonnes, which is above the long-term (1946-2008) average. The sequenced cod (NEAC_001) was a wild-caught male from the NEAC population, and estimated at 8 years of age based on otolith rings. Due to the large population size we expected NEAC_001 to exhibit substantial levels of heterozygosity. Additionally, we used genomic resources from a coastal cod specimen (NCC_001) for which a BAC library was previously created (Supplementary Note 3).

Supplementary Note 2 DNA isolation

High molecular weight DNA agarose blood plugs were made according to the supplementary protocol of Oesagawa et al. (2001) from the NEAC_001 and stored in 0.5 M ethylenediaminetetraacetic acid (EDTA). DNA was dissolved overnight in 1 ml of TE-buffer. Quality and quantity of DNA were checked using NanoDrop (NanoDrop Products), PicoGreen Quant-iT™ (Invitrogen) and FLUOstar Optima (BMG Labtech) and through visual inspection of agarose gels.

Supplementary Note 3 Shotgun and paired-end libraries

The 454 data set was obtained by sequencing a combination of shotgun and paired end libraries from NEAC_001 using the GS FLX Titanium chemistry. The libraries were constructed using the Roche kits and protocols at the Norwegian Sequencing Centre (www.sequencing.uio.no) and 454 Life Sciences, Branford, USA. The total read data set consisted of 63.6 million shotgun reads (peak read length 503 bases, Supplementary Figure 1) and 20.2 million paired-end reads (peak read length 389 bases, including the linker sequence; average pair half length 81 bases, Supplementary Figure 1), with jumping distances of 1 to 2Kb, 3Kb, 8Kb and 20Kb (Supplementary Table 1).

An Illumina sequencing library was prepared (fragment length ~300 bp) from NEAC_001 DNA and a full run on the GaIIx of 2x76 bp was done, according to the Illumina protocols at the Norwegian Sequencing Centre. In this way, 89 million paired reads were obtained, totalling 13.5 Gbp of raw sequence.

A BAC library (insert size between 115Kb and 170Kb in pECBAC1) was constructed from NCC_001 sperm (Amplicon Express). A total of 91,195 end-sequences, of which 78,034 were pairs from 39,017 BACs, were obtained as described in Kuhl et al. (2011) using Sanger sequencing.

Supplementary Note 4 Genome size estimate

Based on analysis of haploid DNA content or C-values the Atlantic cod genome size was estimated to be 930 MB (Hardie and Hebert 2003, Hardie and Hebert 2004). Nevertheless, C-value estimates can vary and a substantially lower genome size of 420Mb was also suggested (Grosvik and Raae 1992).

We calculated genome size based on average sequencing depth, as determined by the peak in the frequency distribution of unique k-mers in the total raw sequencing read data set (Li R. et al. 2010). All shotgun and paired-end sequencing reads were used for this analysis and quality trimmed using the –tr option in Newbler to exclude low-quality bases. The profiles of the k-mer distributions were obtained using meryl (http://sourceforge.net/apps/mediawiki/kmer/index.php?title=Main_Page). After determining the distribution profile, the estimated sequencing depth can be calculated by the formula $E = D * (L – k + 1) /L$, where D is the sequencing depth, L is the average read length, k is the k-mer length, and E is the peak depth (mode) obtained from the profile (Supplementary Figure 2). Total size can be calculated by dividing the total amount of sequenced bases by sequencing depth. The length of the k-mers influences the final genome size estimate as the peak depth decreases with an increase in k-mer length. Nevertheless, peak depth did not decrease for k-mer sizes longer than 20 bases in length and for which a well-defined peak was present. Thus, a k-mer length of 20 bases was chosen - giving genome size of 830Mb (Supplementary Figure 2), slightly lower than the estimates based on haploid DNA content.

Supplementary Note 5 Newbler assembly

Before assembly of the 454 reads, we excluded highly repetitive, non-informative reads from our data set. Shotgun reads that consisted entirely of short tandem repeats (STRs, mono-, di-, tri- and tetranucleotide repeats, allowing a maximum of 10 non-repeat bases at the beginning or end of the read) were excluded. Using the same criteria, sets of paired reads were excluded if at least one of the pair halves consisted entirely of such repeats. Assemblies based on subsets of the read data set with or without such highly repetitive reads showed comparable assembly statistics, but calculation time and computer memory usage was substantially improved when the repeat reads were not included (data not shown). The final read data set was assembled using Newbler (Margulies et al. 2005) version 2.3 (PostRelease-11/24/2009) with the ‘large’ flag (for large genome assemblies) set, using 24 CPUs and a maximum of 78 Gbytes of memory. The assembly (ATLCOD1A) consisted of 6,467 scaffolds totalling 611Mb (a scaffold is here defined as at least two contigs connected with at least two consistent paired end reads). There were 143,207 gaps in the scaffolds (226 Mb, average gap size 1575 bp).

Contigs that could not be assembled in scaffolds had a read depth (the number of bases from all the reads used to assemble the contig, divided by the contig consensus length) of approximately half of scaffolded contigs, despite having a similar GC% content (Supplementary Figure 3). The read depth of these unscaffolded contigs suggests that these represent heterozygous portions of the genome. Polymorphic regions of the genome were apparently assembled into two separate contigs, which subsequently showed approximately half the read depth of that of the homozygous portions of the genome. The peak read depth of the scaffolded contigs was 40.8x and

indicates the sequencing read depth of the homozygous single copy regions of the genome.
Paired end reads were mapped back to the contigs of the assembly and distances between the pair halves ('mates') determined (Supplementary Figure 4). The 20Kb libraries showed an extra peak around 3Kb; as a result of the small number of contigs of at least 20Kb, some reads were artificially constrained to map to shorter contigs.

Supplementary Note 6 Celera assembly

Among alternative assembly programs tested, only the Celera Assembler produced results of similar quality to Newbler. The Celera assembler from CVS (dated 13.7.9, http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Check_out_and_Compile) was used to assemble the 454 read data set, and in addition the 39,017 NCC_001 BAC paired end reads (sequenced using Sanger, see Supplementary Note 3). The 454 SFF files were converted to FRG file format required by Celera assembler using *sffToCA*, and filtered for lengths below 200 and above 800 bases, and overly repetitive reads, using *remove_fragment*. The Sanger BAC reads were converted to FRG using *convert-fasta-to-v2.pl*. Various conditions were tested to optimize the assembly using strategies similar to those documented in the Celera assembler Wiki (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Consensus_Failure).
We experimented with different parameters and found the following to work well: utgErrorRate = 0.03; ovlErrorRate = 0.06; overlapper = ovl; unitigger = bog; ovlMemory = 8GB --hashload 0.8 --hashstrings 400000; ovlHashBlockSize = 400000; ovlRefBlockSize = 4000000; ovlConcurrency = 16; ovlThreads = 2; ovlCorrConcurrency = 16; frgCorrThreads = 16; frgCorrConcurrency = 16; merOverlapperSeedConcurrency = 16; merOverlapperExtendConcurrency = 16.
The final Celera assembly (ATLCOD1B) of ~45M reads has a total size of 629 Mb and N50 scaffold size of 488,312 Kb (Table 1, main text), with 127,414 gaps (71 Mb, average gap size 559 bp).

Supplementary Note 7 Short tandem repeats

Using Tandem Repeat Finder (TRF, http://tandem.bu.edu/trf/trf.html, Benson 1999), all 302,419 contigs of at least 500 bp were screened for short tandem repeats (STRs) of 1-4 bases in repeat length (homopolymers, di-, tri- and tetranucleotide repeats). The distance from the beginning of the repeat to the beginning of the contig, or from the end of the repeat to the end of the contig (whichever was the shortest distance) was determined and the cumulative percentage of contig edges that have an STR starting at that distance was plotted (Supplementary Figure 5). Close to 32% of the contig edges started with an STR, while 38% of the contig ends had an STR within the first 10 bp. The majority of STRs were 2-mers (dinucleotide repeats).
In the scaffolds, we found 24% of the gaps flanked on both ends by an STR and 11% of the gaps flanked by an identical STR (with the start of the STR within the first 10 bp from the gap beginning and end in both cases). The sizes of these particular gaps were on average shorter (peak length 228 bp) compared to those of all gaps (peak length 362 bp). The association of gaps with STRs indicate complications when sequencing and assembling STRs using the 454 technology which contributed to the fragmented nature of the Newbler assembly.

Supplementary Note 8 Closing gaps caused by heterozygosity

Newbler produces a 'contig graph' formed by all read alignments, with contigs as nodes, and reads aligned in more than one contig as edges (Quinn et al. 2008). Alternate paths through the graph that have their start at the same node, and converge on another node, and for which the contigs in between show sequence similarity are often referred to as 'bubbles' in the graph. Such paths are indicative of small sequence variation caused by sequence errors or biologically 'real' sequence polymorphisms (Zerbino and Birney 2008). Newbler resolves such bubbles by introducing gaps in the corresponding scaffolds, which results in fragmentation of the contigs.

Using a custom script, the contig graph was traversed in order to detect the presence of gaps formed by such bubbles. For those gaps where the alternative paths spanning the gap consisted of up to two separate contigs, the path with the highest number of reads going from the 5' end of the flanking contig into the gap was chosen. This chosen path was then used to connect the contigs flanking the gap, and both paths were reported as the alternatives for a polymorphic region. In some cases, one of the paths was a direct path (no contigs) between the two flanking contigs, while the alternative had one or two contigs spanning the gap. In those cases, the polymorphism was an insertion/deletion. The script also detected erroneous gaps in scaffolds, i.e. some gaps were reported despite a continuous assembly contig graph. Overall, this 'heterozygote gap closure process' eliminated 15,765 gaps (11%) in the reference sequence and indicated 4,582 polymorphic regions. By concatenating the newly connected contigs (without gaps) into a single, longer contig, the assembly contig N50 of the *scaffolded* contigs increased to 3,943 bp. 582 scaffolds were concatenated into a single contig and these regions are not reported as scaffolds in the final assembly metrics unless polymorphic information was present for the original gapped region. Overall, after the gap closure process, the number of scaffolds was reduced to 6,467, scaffold N50 increased to 688Kb and total assembly length became 611Mb, see Table 1. There were 134 scaffolds (2% of the total) of at least 1Mb, totalling 236 Mb (39% of the total scaffold length).

Supplementary Note 9 Comparing the Newbler and Celera assemblies

The size distributions of the Newbler and Celera contigs and scaffolds were compared by plotting the proportion of the assembly present in those of a certain minimum size (Supplementary Figure 6). The Celera assembly had a larger proportion of the assembly in contigs above 10Kb, while the Newbler assembly showed a larger proportion of the assembly in scaffolds over 1Mb.

To compare the Newbler and Celera assemblies, we used ATAC (Istrail et al. 2004). This program starts with a seed alignment and extends it as long as there is 100% identity. Such alignments are appropriately merged to get so-called 'clumps' while maintaining > 95% identity. The parameters are optimized so that self-comparison of the assemblies returns same scaffolds. The reported clumps were parsed to obtain total length matches between scaffolds from the assemblies. These were sorted in descending order and represented as a density plot (Supplementary Figure 7) for the top 400 Celera scaffolds (representing 284Mb of the assembly) and the top 300 Newbler scaffolds (330Mb), totalling a match of 226Mb between them. There was an overall trend of >90% agreement between these assemblies, with rearrangements depicted by the off diagonal points.

Supplementary Note 10 Metrics of Fish Genomes

When comparing the assembly metrics to those of other fish genome assemblies (Aparicio et al. 2002, Jaillon et al. 2004, Kasahara et al. 2007, http://www.ensembl.org/Gasterosteus_aculeatus, http://www.ensembl.org/Danio_rerio), including previous versions, the Celera assembly showed comparable contig metrics with respect to medaka, with the Newbler assembly having both more contigs and a slightly higher contig N50 number (Supplementary Figure 8). The Newbler scaffolds showed a comparable scaffold N50 number and scaffold count as the zebrafish, fugu and medaka scaffolds (Supplementary Figure 8).

Supplementary Note 11 Heterozygous SNPs

All available 454 and Illumina reads (Supplementary Note 3) were mapped to the repeat-masked contigs (Supplementary Note 16). For the 454 reads, the Newbler runMapping command was used with overlap length of minimum 80% of the read length, and minimum 96% identity in the overlap (settings: -ml 80% -mi 96). BWA (Li H. and Durbin 2009) was used for the Illumina reads with default settings.
The candidate SNP set was filtered according to the following thresholds: 1) At least 3 reads should share the polymorphism; 2) No other SNPs should be detected in a 5 base pair window on either side of the SNP. Between Newbler (454 reads, 603,555 SNPs) and BWA (Illumina reads, 873,847 SNPs), there were 429,527 SNPs in common. This amounted to a total of 1,047,875 SNPs (603,555+ 873,847 - 429,527), indicating a heterozygosity rate of 2 SNPs/Kb (1,047,875 SNPs; 500,614Kb repeat-masked genome sequence).

Supplementary Note 12 RNA isolation and preparation and sequencing of cDNA

We prepared cDNA libraries from a number of tissues of NEAC_001 (Supplementary Table 2), and sequenced them using 454 GS FLX (Titanium chemistry). Isolation of total RNA was done with the RNeasy Midi Kit (Qiagen), using 70 – 100 mg tissue (without addition of β-mercaptoethanol during the RNA cleanup step). Poly-A+ RNA was isolated with the Oligotex mRNA mini kit (Qiagen). First strand cDNA synthesis was performed with RevertAid™ H Minus First Strand cDNA Synthesis Kit (Fermentas) where the synthesis time was reduced from 1 hour to 7.5 minutes to obtain cDNA of the desired length for 454 library preparation. Second strand synthesis was performed with DNA polymerase I, *E.coli* (Fermentas) according to standard protocol. The double stranded cDNA solutions were cleaned up with MinElute PCR Purification Kit (Qiagen). Libraries for 454 sequencing were produced and sequenced following the Roche protocol for shotgun Titanium standard library preparation. A total of 1.4 million 454 Titanium reads were obtained (Supplementary Table 2).

Supplementary Note 13 Assembly and mapping of the Atlantic cod transcriptome

We added available cDNA data from previous sequencing efforts (Johansen et al. 2009, the Cod Genomics and Broodstock Development (Canada, unpublished data, http://codgene.ca), Mari Moren et al (unpublished data), Lie et al. 2009, Olsvik and

Holen 2009, Edvardsen et al. 2010) (Supplementary Table 2) to our total cDNA dataset (Supplementary Note 12), leading to a total of 1.6 million reads.
The cod transcriptome was assembled using Newbler (v2.3, options/settings -cdna -large -ace -mi 98 -ml 90, with filtering of the reads against cod 18S, 5S, 28S ribosomal and mitochondrial genes) leading to 41,419 contigs after quality trimming (as described for reads above), with a peak read depth at 3-4x (Supplementary Table 3). In order to obtain contigs for the most highly expressed transcripts, we assembled a random subset (29%) of the reads labelled 'repeat' in the assembly, using the same parameters, adding 149 new contigs. Using BLASTn (Supplementary Table 5). 99.5% of the 41,568 contigs could be aligned to the Newbler contigs and 94% to the Newbler scaffolds (maximum e-value 10^{-9}). The mapping to the Celera assembly showed slightly lower number of aligned contigs, with more directional and fewer positional alignment errors (Supplementary Table 4), compared to the mapping to the Newbler assembly.
All transcriptome contigs were annotated using the GAFFA CDS Prediction Pipeline (http://genofisk.cbu.uib.no; unpublished) developed at the University of Bergen in collaboration with the National Norwegian Institute for Marine Research. In total, 15,068 coding sequences were predicted (4,857 with full length CDS). Putative frameshifts were identified in 4,450 transcriptome contigs relative to the best protein match and premature stop codons in another 2,271, possibly related to low sequencing depth, homopolymer stretches and problems with accurately predicting intron boundaries in predicted transcripts.

Supplementary Note 14 BAC insert Sequencing

Based on mapped BAC ends (Supplementary Note 15), two sets of BACs were selected so that i) each pair mapped to two different regions of the genome, ii) there was a significant overlap between the BACs of each pair. Shotgun sequencing of selected BAC clones was performed according to Negrisolo et al. (2010). The sanger reads (AB1/SCF format) were converted to fasta using phred (http://www.phrap.org/phredphrapconsed.html) and further converted to frg using convert-fasta-to-v2.pl (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Formatting_Inputs#convert-fasta-to-v2.pl). Assembly of each individual BAC data set was done using the Celera assembler (CVS tip Feb 2010, http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Check_out_and_Compile) using the following settings: utgErrorRate = 0.015, ovlErrorRate = 0.06, overlapper = ovl, unitigger = bog, ovlMemory = 2GB, ovlHashBlockSize = 200000, ovlRefBlockSize = 2000000, ovlConcurrency = 1, ovlThreads = 2, ovlCorrConcurrency = 4, frgCorrThreads = 2, frgCorrConcurrency = 8, merOverlapperSeedConcurrency = 1, merOverlapperExtendConcurrency = 1.
The best assembly was for BAC 24g13 (two scaffolds), the other BACs had between 6 and 20 scaffolds (Supplementary Table 5). The BAC scaffolds were aligned to the Newbler and Celera scaffolds that the BAC regions covered using NUCmer (Kurtz et al. 2004) version 3.07 (with default settings). Mummerplot (version 3.5) was used with the NUCmer delta file as input (with settings -f so that only alignments which represented the "best" one-to-one mapping were displayed, and -l to order and orient the alignments such that the largest hits clustered near the main diagonal, Supplementary Figure 9).
The BAC scaffolds of BACs 24g13, 45a16 and 62j04 were put in the same order and orientation when Newbler or Celera was used for ordering, and showed no apparent

misassemblies for BACs 24g13 and 45a16.The alignments of BAC 29j05 to Newbler scaffold 2025, and of BAC 62j04 against Celera scaffold 1551617, showed potential assembly errors (split in the diagonal plot). The alignment of BAC 29j05 against the Celera scaffold showed many contigs not aligned; these aligned instead in the middle of another Celera scaffold (scaffold 155123, not shown), indicating a misassembly in the Celera scaffolds. The ordering of BAC 29j05 scaffolds was different when aligned to the Newbler and Celera scaffolds.
The WGS assemblies of the BAC inserts are highly consistent with the Sanger-based assemblies and show limited rearrangements. In fact, the 454 read data set yielded better assemblies of these genomic regions than the individual Sanger sequenced BACs.

Supplementary Note 15 BAC end mapping

The 39,017 NCC_001 BAC paired end sequence reads were masked using a Atlantic cod specific repeat library (Supplementary Note 16) and subsequently mapped to the 6,467 masked scaffolds from Newbler assembly and 17,338 scaffolds from Celera assembly using the gsMapper software from the Newbler package (minimum overlap length 40% of the read length, minimum 90% identity in the overlap). 10,171 read pairs were mapped on the same Newbler scaffold and 14,618 pairs on the same Celera scaffold with a peak distance of 94,465 and 97,871 bp respectively (Supplementary Figure 10). Since the Celera assembly included the BAC end reads, we expected more pairs mapped to be mapped to this assembly. The small peak around very low mapping distances may be due to BACs with a very short insert size, as BAC libraries typically contain a fraction of small clones.
The BAC end mapping results indicate long-range contiguity of the scaffolds for both assemblies, even though the BAC-ends were not included in the Newbler assembly.

Supplementary Note 16 Repeat Analysis

Known repetitive and transposable elements (TE) were identified and masked using RepeatMasker (Version 3.2.8, Smit et al.) with the RepBase Update (RU) TE library (RM database version 20090604) (Jurka et al. 2005). Using this approach, 12.31% of the assembled genome was masked (Supplementary Table 6). Subsequently, we created a custom *de novo* repeat library by identifying novel TE elements in the Atlantic cod genome using RepeatModeler (Smit and Hubley). RepeatModeler uses the programs RECON(Bao and Eddy 2002), RepeatScout (Price et al. 2005) and Tandem Repeat Finder (Benson 1999) to generate families of TEs. Additionally, this method aims to classify these families based on sequence similarity to known TEs. RepeatModeler identified 1124 novel repeat families of which 335 were classified into their respective TE type (e.g. DNA, SINE, LINE or LTR).
In an attempt to classify a larger fraction of the novel repeats, the repeat families were further analyzed with TEclass (Abrusan et al. 2009). This approach classifies the repeats into their main taxonomic branches, reflecting the mechanism of transposition. TEclass classified 839 of the novel repeat families. A total of 305 families were classified twice, both with RepeatModeler and TEClass. Of these, 164 families were classified identically. We used the classification based on sequence similarity (RepeatModeler) whenever there was a difference in classification between these approaches. Despite the increased level of classification of the TE using TEclass, the *de novo* identification and characterization of repetitive elements is known to be notoriously difficult (Bergman and Quesneville 2007, Lerat 2009) and the repeat library likely represents an underestimated proportion of families of TEs in

the Atlantic cod genome. The *de novo* repeat library was compared to UniProt (release 15.13) using BLASTx to remove any coding sequences that were classified as repetitive element.
Results from RepeatMasker, the *de novo* repeat library and the teleost repeat library from RepBase were combined to mask 25.4% of the assembled Newbler genome (Supplementary Table 6).

Supplementary Note 17 Annotation/Gene content

An initial assessment of annotations generated by the standard Ensembl pipeline on the Newbler assembly revealed that a high proportion of the annotations were fragmented due to the fragmented nature of the genome assembly. Therefore, the Atlantic cod protein-coding genes were annotated using both the standard Ensembl gene annotation pipeline and the Ensembl gene projection pipeline. The Ensembl gene projection pipeline has been used previously by Ensembl to project protein-coding annotation from the human genome onto high-coverage primates and various mammalian low-coverage genomes, which have been sequenced as part of the Mammalian Genome Project; however, this approach has not been used before to project annotation between species of the Actinopterygii class.
The three-spined stickleback, *Gasterosteus aculeatus,* was chosen as the reference species from which gene annotations would be projected onto cod. The stickleback is a well-characterized fish species annotated by Ensembl and the genome has a base coverage of 11x. The stickleback annotations from Ensembl comprise 20,787 protein-coding genes.
The initial stage of the Ensembl gene projection pipeline involved a whole-genome alignment between the genomic sequence of Stickleback and Atlantic cod. The resulting alignments were filtered to include only genomic regions where the two species had a very close sequence match and where the alignments were reciprocal best-in-genome. Next, sequence regions of the cod genome were reordered and joined into 'gene-scaffolds' based on the whole-genome alignment information and guided by the gene structures on the stickleback genome. In the final stage of the projection pipeline, the protein-coding gene models from stickleback were projected through the whole-genome alignments onto the re-ordered cod genomic sequence regions ('gene-scaffolds', ATLCOD1C).
Major benefits of this projection annotation strategy are the integration of previously unplaced contigs into existing cod genome scaffolds to reflect shared synteny between cod and stickleback. The projection of gene structures onto the newly created 'gene-scaffold' sequence regions resulted in longer, less fragmented gene models. Furthermore, missing exonic sequence was indicated by 'gap exons' in sequence gaps to reflect the gene structure of the reference species and to avoid frameshifts.
With the projection pipeline, 17,920 of 20,787 stickleback protein-coding genes were projected onto the cod genome sequence. 31,135 previously unplaced contigs, totalling 26 Mb (including gaps), were integrated into the resulting gene-scaffolds.
In genomic regions where no genomic alignments with stickleback were generated, or where the projection method failed for other reasons, additional protein-coding gene models were generated using the standard Ensembl pipeline. Ensembl annotations for stickleback, *Oryzias latipes* HdrR (medaka) and *Danio rerio* zv8 (zebrafish) were mapped to the cod genome with exonerate (Slater and Birney 2005) and UniProt proteins were mapped to the genome with Genewise (Birney et al. 2004). The resulting annotation was scanned for pseudogenes and integrated with the results from

the projection pipeline, along with the results of the Ensembl non-coding RNA pipeline.
The final cod gene set comprises 22,154 genes in total. This includes the 17,920 protein-coding genes identified using the projection approach, 2,175 protein-coding genes annotated using the standard Ensembl pipeline, and a set of 2,059 pseudogenes and short non-coding RNA genes (Supplementary Table 7).
The majority of the annotated 20,095 protein-coding genes matched to the contigs of the transcriptome assembly (72%, BLASTn, maximum e-value 10^{-9}, Supplementary Note 13). Conversely, 13,941 of the 15,068 predicted CDSs in the transcriptome contigs were mapped to the protein-coding gene models (92.5%, BLASTp, maximum e-value 10^{-9}). The transcriptome data set does not contain all annotated genes however, this dataset is well represented in the predicted gene set.

Supplementary Note 18 GO terms associated with Atlantic cod genes

GO terms were extracted from Ensembl databases (release 57) for the sequenced fishes (stickleback, tetraodon, fugu, medaka and zebrafish), and the Atlantic cod genome annotation. For Atlantic cod, 12,967 genes had 42,487 GO terms (1625 unique terms) associated with them. The online tool CateGOrizer (Hu et al. 2008) was used to cluster terms according to the GO Slim2 classification method (single occurrence counting). Compared to other sequenced teleosts, Atlantic cod GO terms have a similar distribution, though terms for catalytic activity and metabolism are overrepresented, while terms associated with development are underrepresented (Supplementary Figure 11).

Supplementary Note 19 Linkage map SNPs

The flanking sequences of the SNPs used to construct the Atlantic cod linkage map (Hubert et al. 2010) were mapped to the Newbler and Celera assemblies using BLASTn (word size 25, maximum e-value 0.01). For each SNP with linkage information, the best hit was used to associate this information to the corresponding scaffolds (and some contigs not in scaffolds). For each assembly, different numbers of scaffolds were associated with the linkage map (Supplementary Table 8) The available linkage map for Atlantic cod is not dense enough to order the scaffolds along the linkage groups. Yet, in several cases, scaffolds were mapped to the linkage map through multiple SNPs. For the majority of cases, and in particular when comparing scaffolds assembled by Newbler, these scaffolds are then associated multiple times to the same linkage group, as expected. Only in a few cases are scaffolds associated with multiple linkage groups, thus contesting the linkage group evidence. These contesting results may occur due to assembly or linkage map errors. Despite the low density of SNPs in the linkage map, 332Mb and 274Mb of the Newbler and Celera assemblies, respectively, is associated with the 23 linkage groups. The Newbler assembly has more sequence anchored and has the highest percentage of possible comparisons consistent with linkage group evidence (98.3%).

Supplementary Note 20 Synteny

We used the Ensembl release 57 of stickleback (*Gasterosteus aculeatus*), tetraodon (*Tetraodon nigroviridis*), fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), zebrafish (*Danio rerio*) and human (*Homo sapiens*) genome assemblies for the synteny analysis.

The locations of the longest annotated Atlantic cod transcripts (20,095) from the reorganized gene-scaffolds (Supplementary Note 17) were translated to the original Newbler assembly. These sequences were then located in the zebrafish, tetraodon, stickleback, medaka and human genomes by aligning the translated protein sequences using gmap (Wu and Watanabe 2005). All alignments with at least 50% of the transcript and at least 50% sequence identity were considered orthologous. In those cases where multiple alignments were found for a gene, the match with highest sequence alignment scores was selected. The locations of orthologous genes were then pair-wise compared to those in Atlantic cod for each species. For the fraction of the genome that could be assigned to the linkage map (consisting of approximately 332Mb, Supplementary Note 19), we counted the co-occurrence of the orthologs among the linkage groups of Atlantic cod with those located on the chromosomes of the other teleosts, and human for reference (Figure 1, Supplementary Table 9). Of the four teleosts, stickleback, tetraodon and medaka have the highest number of genes assigned to particular linkage groups. This pattern is also present in a similar analysis incorporating the entire Atlantic cod sequence (data not shown).

Supplementary Note 21 Haemoglobin

We previously resolved the structure of the haemoglobin cluster by sequencing BAC clones from a Norwegian coastal cod (NCC) individual, designated NCC_001, (Wetten et al. 2010). By comparing the haemoglobin clusters in the genome (specimen NEAC_001) to the independently assembled BAC insert sequences we found an insert of 73 bp in the intergenic promoter region (Figure 2a, main text). Detailed alignments are available on request.

Supplementary Note 22 Linkage disequilibrium in the *ß1* globin region

The extent of linkage disequilibrium between the polymorphic *ß1* promoter and the polymorphic sites in *ß1* globin was evaluated for all three locus pairs (promoter *versus ß1*-55, promoter *versus ß1*-62 and *ß1*-55 *versus ß1*-62) across all eight populations (Supplementary Figure 12, Supplementary Table 10) using Fisher's method through Genepop (http://genepop.curtin.edu.au/, Rousset 2008). All three comparisons revealed significant linkage disequilibrium (Chi^2=infinity, df 16, p-value=0) among the different alleles.

Supplementary Note 23 Transfection experiments

The two *α1-ß1* globin promoter variants were obtained by direct PCR amplification of the complete intergenic region between the start codons of the 5'-5'oriented *α1* and *ß1* globin genes using genomic DNA from specimens homozygous for each of the two promoter variants (short and long) as template. A set of sense/antisense primers targeting both promoter variants were designed with restriction sites for XhoI and HindIII in their 5' ends. A degenerate site (W) in the Hb-HindIII primer reflects a SNP at this locus. Primer sequences are as follows:
Hb-XhoI: 5'-ATCTCGAGCTTGAATAGTGTGGTCAGATTGGACTCTGT-3' and
Hb-HindIII: 5'-ATAAGCTTTGTGGCGTWGTCTTAAGGGTTCAATGT-3'.
The PCR products were sequenced, and plasmids for transfection were constructed by cloning the different promoters into XhoI and HindIII restriction sites in Promega's expression vector pGL4.20, which harbors a firefly luciferase gene downstream of the cloning sites. Vector inserts were sequenced to confirm the promoter types and correct orientation, i.e., the promoter end originating from upstream the *ß1* globin start-codon oriented upstream of the luciferase reporter in the expression vector. An

internal control vector for co-transfection (Promega pGL4.73) harboring SV40 early enhancer/promoter and renilla luciferase reporter gene was used in the experiment without modifications.
TO-cells from Atlantic salmon (Wergeland and Jakobsen 2001) were cultured in L-15 medium (Cambrex Bio Sciences) supplemented with 50 μg/ml gentamicin, 4 mM L-glutamine, 40 μM β-mercaptoethanol and 10% fetal calf serum. After splitting, the cells were adapted to three different temperatures (4, 15 and 20 °C) for one week in 175 cm^2 flasks before trypsinization and transfection. One μg of globin promoter construct and 1 μg of internal control construct were co-transfected into ~$2*10^6$ cells using an AMAXA Nucleofector™ device (program T-20) according to the manufacturer's instructions (Lonza AG). Transfection efficiency was routinely about 50-60 % (assessed by parallel GFP transfection). Three transfections were performed per promoter variant and temperature, and directly after transfection the cells from each transfection were suspended in growth medium and seeded in triplicate to 6-well plates before incubation for another week at same temperature as prior to transfection. Effects of temperature on expression efficiency were determined by analyzing firefly and renilla luciferase activity using the Dual-Luciferase® Reporter Assay System according to the manufacturer's description (Promega). Relative expression efficiency for the promoters were obtained as ratios of firefly luminescence (controlled by the Hb promoters) to renilla luciferase expression from the pGL 4.73 internal control vector to eliminate bias resulting from possible differences in transfection efficiency, unequal cell numbers and other per sample related errors. Luminescence ratios were calculated against a co-transfected SV40/Renilla reporter and normalized for each separate transfection experiment. The ratios were normalized to allow comparison between the three separate transfection experiments. The data were analyzed using a GLM (Generalized Linear Model), with temperature and promoter type as crossed factors, the three replicates as nested factor in both temperature and promoter type, and normalized expression ratio as response. The interaction between temperature and promoter type was highly significant (GLM, $F_{2,36} = 7.85$ $P = 0.007$, Figure 2b, main text).

Supplementary Note 24 Pairwise comparison of identity scores of human immune genes among teleosts

We aligned 1308 human genes involved in immune response (associated with the GO term GO:002376 "immune system process", http://www.geneontology.org) to the genomes of Atlantic cod, fugu, tetraodon, medaka, stickleback, and zebrafish using exonerate (Slater and Birney 2005). For each protein the identity scores of the best matches were plotted in a full pair-wise comparison (Supplementary Figure 13). No particular pattern is present that sets Atlantic cod apart from the other teleosts.

Supplementary Note 25 Investigation of unassembled, unfiltered sequencing reads for the presence of MHCII, CD4 and the invariant chain

The presence of traces of MHCII, CD4 and the invariant chain (CD74) genes in Atlantic cod was progressively investigated by querying a diverse set of vertebrate homologs (Supplementary Table 12) to the genome assemblies, cDNA assembly, and eventually the original unassembled 454 and Illumina sequence reads. The unassembled sequence reads represent a dataset of ~49.5 Gb in size, with ~36 Gb generated by the 454 platform and ~13.5 Gb generated by the Illumina platform through the creation of multiple independent libraries (Supplementary Table 1, Supplementary Note 3). For CD4 we find partial evidence in both assemblies. We

located a fragment on contig102546 in the Newbler assembly (Supplementary Table 11) and on scf7180001550564 in the Celera assembly (Supplementary Figure 18, Supplementary Note 27). The region where the CD4 fragment is located is over 99.92% identical in both assemblies and no other CD4 homologs or fragments were found in either assembly.

For MHCII and the invariant chain we obtained no reciprocal blast results, unambiguously locating homologs or fragments, in the assemblies, cDNA assemblies or any unassembled sequence trace generated by the sequencing platforms. Reciprocal BLAST results were obtained by searching the NCBI RefSeq database (Release 42). This complete absence is striking: The MHCII and invariant chains genes are usually present in several copies located on different chromosomes or chromosomal regions in the other sequenced teleost genomes (Supplementary Note 27). The genomic distribution of these genes in other species implies that their loss in Atlantic cod has evolved through several independent events.

Supplementary Note 26 qPCR targeting teleost MHCII

We designed primers to amplify conserved regions in the exon 3 of the MHCII beta chain gene. Conserved regions were located in the sequences of a selection of teleost species (Supplementary Table 13) that represent a phylogenetically diverse group, yet show sufficient sequence similarity to enable primer design (Supplementary Figure 14). Based on the alignment of these sequences, 10 primers were designed targeting two regions. Several primers contained degenerate sites (Supplementary Table 14). All primers were tested on the LightCycler480 instrument, following standard protocol. In stickleback, two primer pairs (Class_IIB_ex3_Forw1 + Class_IIB_ex3_Rev1.1, and Class_IIB_ex3_Forw1 + Class_IIB_ex3_Rev1.3) provided the best amplification and resulted in an amplicon size of 116bp and 161bp respectively (data not shown). In Atlantic cod, all primers pairs performed similarly inadequate. Using the two best performing primer pairs, a dilution series (undiluted, 10X, 100X, 1000X, 10000X) was additionally run in three replicates on Atlantic cod and stickleback. We simultaneously amplified a region of the Beta-2-microglobulin (Supplementary Table 14) with species-specific primers as a positive control, following identical methods and primer design as described in the MHCI copy number estimation (Supplementary Note 28). All reactions were run on a single 96 wells plate.

In stickleback, the qPCRs that are combination of dilutions and replicates of the two primer pairs targeting MHCII resulted in detectable fluorescence levels well before those of the negative controls (Supplementary Figure 15a). These MHCII levels appear slightly later than those of the positive controls targeting the Beta-2-microglobulin, likely due to the fact that the positive control primers are stickleback specific. The amplicon melting curves for these reactions show two unimodal distributions (Supplementary Figure 15b). The melting curves of the positive controls are located between the two MHCII primer curves. This location is as expected considering the size of the amplicons (116bp and 161bp for the MHCII regions, and 129bp for the positive controls).

In Atlantic cod, the primer pairs targeting the conserved regions in MHCII do not reach detectable fluorescence levels before those of the negative controls (Supplementary Figure 15c). The amplicon melting curves of the primer pairs have multivariate distributions, which are dominated by those consisting of primer – dimers, similar to those of the negative control (Supplementary Figure 15d). Thus, in Atlantic cod, both the appearance of detectable levels of fluoresence and the melting

curves are not distinguished from the background levels present in negative controls. Instead, the positive controls targeting the Beta-2-microglobulin in Atlantic cod result in early detection of fluorescence levels and a clear unimodel distribution of expected size.

Supplementary Note 27 Syntenic regions around MHCII, CD4 and Invariant chain (CD74) among six sequenced teleosts

A selection of well-annotated, protein-coding genes surrounding MHCII, CD4 and the invariant chain was used to study the syntenic landscape around these genes in the sequenced teleosts and Atlantic cod. Copies of homologs for MHCII, CD4 and the invariant chain genes occur on different chromosomes and regions in the teleosts genomes.

For MHCII, the teleost genomes show evidence for at least two sets of genes, with each set containing a gene for the MHCII alpha and beta chain. The zebrafish genome contains most copies of MHCII with five sets of both MHCII genes, organized in two regions (Supplementary Figure 16). The teleost genomes exhibit limited evidence of conserved synteny around the MHCII region. The causes for this absence of evidence are twofold. First, in several teleosts, some of the MHCII genes are located in regions that are not well resolved in their respective assembly and are placed on relatively small, unconnected scaffolds. For example, in fugu (scaffold_7721, scaffold_8524), tetraodon (scaffold7373) and medaka (scaffold837), several MHCII genes have no associated flanking genes. Second, the flanking genes of MHCII genes that are placed in more resolved regions are not syntenous among teleosts. For example, the flanking regions of MHCII in stickleback (chromosome VII), do not agree with those regions surrounding MHCII in zebrafish (chromosome 8) and medaka (chromosome 3).

Because of this lack of overall synteny of the MHCII region in teleosts, we individually compared the order and transcription direction of the genes flanking regions of the species –in those cases where these could be retrieved– to Atlantic cod. The MHCII flanking genes in fugu, tetradon, and medaka are located on several unconnected scaffolds in Atlantic cod, obstructing the identification of a single syntenous region. For stickleback and zebrafish however, we did find substantial evidence for synteny among the flanking genes of their MHCII regions on two scaffolds of Atlantic cod (Supplementary Figure 16 and 17). Both these comparisons show evidence of structural rearrangements in Atlantic cod relative to these two lineages. Within these two syntenous regions in Atlantic cod, we do not find evidence of traces of MHCII.

The structural rearrangements prevent the determination of well-defined boundaries surrounding a potential prior location of the MHCII region in Atlantic cod. Thus, the overall lack in synteny among teleosts and the species-specific rearrangements prevent the detection of a convincing ancestral signature of MHCII genes or their flanking genes in Atlantic cod. Nevertheless, resolving the syntenic landscape of two potential regions in stickleback and zebrafish indicates that at least those corresponding locations are not particularly complex regions that are difficult to assemble in Atlantic cod. The lack of synteny in teleosts around the MHCII region is in contrast to the situation in mammals where a conserved region containing these immune genes is the norm.

The CD4 genes are located in a single region, which exhibits conserved synteny of gene order and transcriptional direction in each of the teleosts genomes (Supplementary Figure 18). Within this region, the number of CD4 genes can range

from one (stickleback) to three (tetraodon). In Atlantic cod, this region did contain a single fragment of the CD4 protein (a truncated pseudogene) in the Celera assembly. This assembly has the CD4 fragment located on a scaffold, whereas the Newbler assembly has placed the same sequence on an unconnected contig (Supplementary Note 25). Apart from the placement of this pseudogene, gene order and orientation of the flanking region is identical in both assemblies. In Atlantic cod, the fragment of CD4 is oriented in the same transcriptional direction as in other teleosts and its open reading frame (ORF) contains a frameshift and two stop codons. We confirmed the location of the CD4 pseudogene independently in the genome by PCR amplification and sequencing (Sanger) of genomic DNA from the sequenced specimen (data not shown).

The CD74 genes occur as two separate homologs, which are located on different chromosomes in each of the teleost genomes. Around one of these homologs we located a region with conserved synteny of gene order and transcriptional direction among all teleosts (Supplementary Figure 19). Interestingly, both the CD74 homologs in zebrafish show synteny towards the same region in the other teleosts. This result indicates that the zebrafish lineage has experienced a specific duplication and deletion of one on the homologs. No evidence for a pseudogenic CD74 sequence could be located in Atlantic cod between the syntenous flanking genes.

The second CD74 homolog shows conserved synteny among fugu, tetraodon, stickleback, medaka and Atlantic cod, but not with zebrafish (Supplementary Figure 20). Both the flanking regions for this second homolog are resolved on two unconnected scaffolds in the Atlantic cod genome assemblies. We did not find evidence for a pseudogenic sequence on either of these scaffolds.

Supplementary Note 28 MHCI copy number estimation

A relative quantification approach based on quantitative real-time PCR (qPCR) was used to estimate the copy number of MHCI in the Atlantic cod genome. We performed the same experiment on the stickleback and human genome to verify the qPCR method.

The α3-domain is the most conserved region of MHCI and therefore most suitable for primer design, and was for that reason chosen to represent MHCI. Beta-2-microglobulin was used as reference single copy gene. One additional Atlantic cod single copy gene, topoisomerase was also analyzed to set a single copy baseline. Atlantic cod (NEAC_001) DNA was isolated from liver using a standard Phenol-Chloroform-Isoamylalcohol protocol, while stickleback DNA was isolated from a fin clip following the protocol of Aljanabi and Martinez (1997). Human DNA was acquired from Roche (catalogue number 11 691 112 001). All primers were designed with Primer3 (Rozen 2000) optimized for Tm $\approx$ 61°C, with a length between 20 and 25nt and a resulting PCR product size of $\approx$ 120bp (Supplementary Table 14 and 15). Single copy genes for Atlantic cod (ENSGAUG00000002225, Beta-2-microglobulin and ENSGAUG00000002208, topoisomerase) were selected based on published information (Persson et al. 1999, Miller et al. 2002), and annotation and read depth of the cod genome. Various BLAST algorithms were used to test all primers against the contigs, scaffolds and raw reads of the Atlantic cod assembly, in order to avoid false positive PCR products. Different dilution series were analyzed for all primer pairs and templates, identifying the window of linearity for each assay. The human and stickleback primers and PCR product were tested in a similar way. We also estimated the number of human and stickleback MHCI loci by *in silico* PCR using primer3 on their respective reference sequences (Supplementary Table 15).

SYBR Green qPCR Master Mix (Roche), with primer concentrations of 0.5μM in a 20μl reaction volume were used for all experiments. Both reference and target (MHCI) were run on the same microtiter plate, with the following protocol; 45 cycles of 95°C 10s, 61°C 10s, 72°C 10s, following an initial 95 °C for 5min denaturizing step on the LightCycler 480 instrument (Roche). The copy number estimations are based on 8-11 separate pairwise comparisons of target and reference PCR assays, each assay analyzed in series of 5 dilutions. The lowest two dilution series were later discarded due to poor reproducibility. Average PCR efficiencies (E) were calculated using Cp estimates provided by the LighCycler 480 software (Version 1.5) for 2nd derivative maximum (Supplementary Figure 21) for each gene individually (Karlen et al. 2007). The following formula was then applied to calculate the diploid ratio between the single copy (sc) reference gene and the multi copy α3 region (Karlen et al. 2007):

$$R_{\alpha 3|sc} = (E^{Cp})_{\alpha 3}/(E^{Cp})_{sc}$$

We obtained bootstrap confidence intervals for all copy number estimates by the random re-sampling of our data sets for all individual genes (Figure 4a). Bootstrap values are based on 50.000 iterations performed using a custom VBA script in MS Excel.

Supplementary Note 29 Phylogeny of MHCI sequences

Primers specific for the amplification of MHCI Atlantic cod cDNA were designed targeting conserved parts of the leader and transmembrane region. These regions were identified using a multiple alignment (ClustalW, Larkin et al. 2007) of all Atlantic cod MHC class I sequences available from NCBI and our cDNA assemblies (Supplementary Note 13). To reduce PCR artefact formation, 8 replicate PCR reactions were run with a reduced number of cycles (25) and longer elongation time (60s) (Lenz and Becker 2008). Traditional TOPO TA cloning was applied, and 384 clones were PCR screened to confirm correct insertion. Of these clones, 192 were Sanger sequenced from both ends. All clones were sequenced at least twice in order to filter out sequencing errors. We removed all duplicate sequences, sequences that were not full-length MHCI cDNA and sequences that contained stop codons or frameshifts in the ORF. Our final data set consisted of 109 different nucleotide sequences, which represented 101 unique amino acid sequences.

The protein alignments of the 101 MHC class I sequences (α1-α3) and those from various teleosts (Supplementary Table 16) were aligned with ClustalW and manually curated in MEGA4 (Version 4.1, Tamura et al. 2007). Tree topology (Figure 4b) and bootstrapping (n=100) for Maximum Likelihood was computed using RAxML (Version 7.2.6, Stamatakis et al. 2005) under the PROTGAMMAIJTTF model, suggested by ProtTest (Version 2.4, Abascal et al. 2005). Bayesian posterior probabilities were calculated using MrBayes (Version 3.1.2, Huelsenbeck et al. 2001, Ronquist and Huelsenbeck 2003), run with 4 chains and with 5.0 million generations, and were sampled every 100th generation. Burnin was set to 40000. Site specific rate model was set to “variable”, and the rate matrix for amino acids set to “fixed (jones)”. Parameters for the likelihood model were set to “invgamma”, and the model allowed the site-specific rate of change to vary over its evolutionary history using the “covarion” setting.

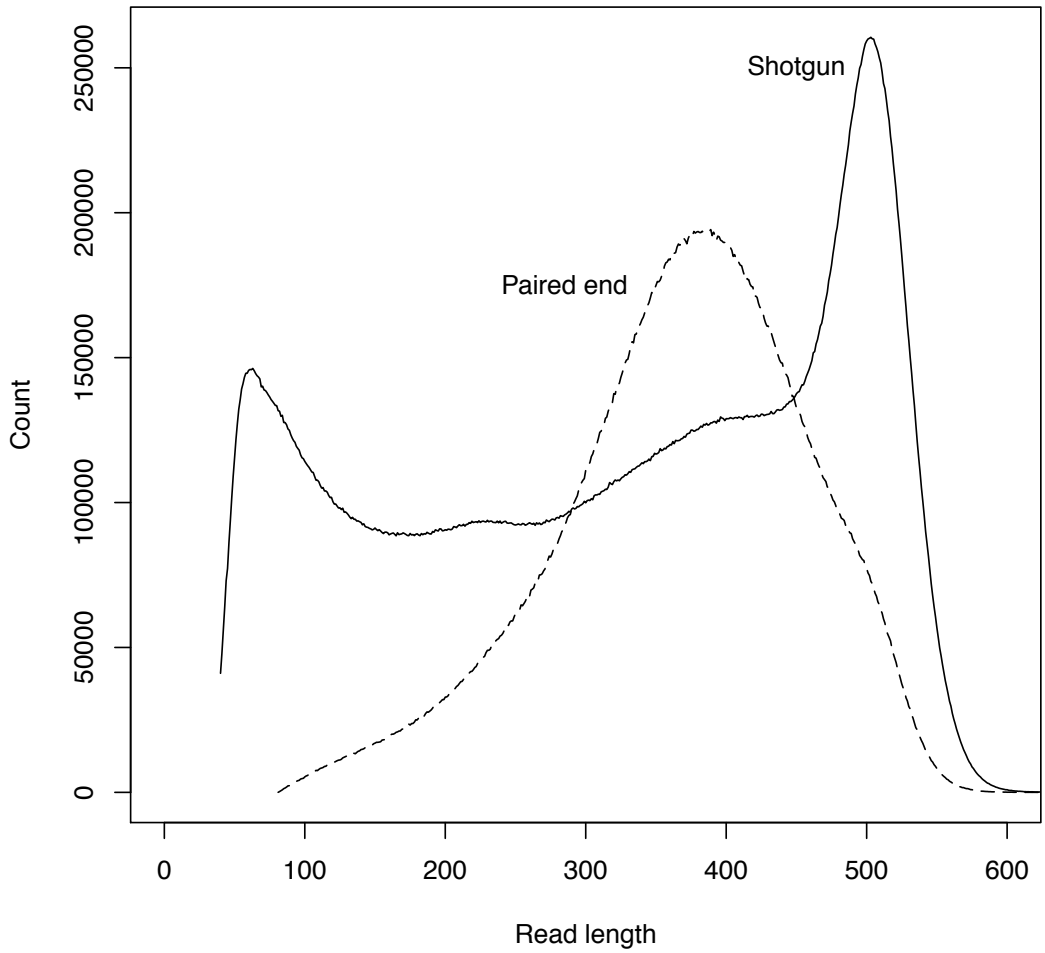
Supplementary Note 30 TLR phylogeny

We obtained Toll-like receptor (TLR) protein sequences from www.ensembl.org for zebrafish, fugu, tetraodon, medaka, stickleback and human. One representative protein sequence was chosen for each TLR (e.g. TRL 1, 2, 3, 4, 5, 6, 7, 8, 9, 14, 18, 19, 20, 21, 22 and 23). TLR11 (rat and mouse specific) and TRL 15/16 (avian specific) were excluded. These TLR sequences were aligned to their respective genomes using the Ensembl webpage BLAST tool in order to retrieve multiple, annotated homologues and TLRs that were not specifically annotated as such in the Ensemble databases. We retrieved all protein homologues with a complete TIR domain, which were used in downstream alignments (Supplementary Table 17).

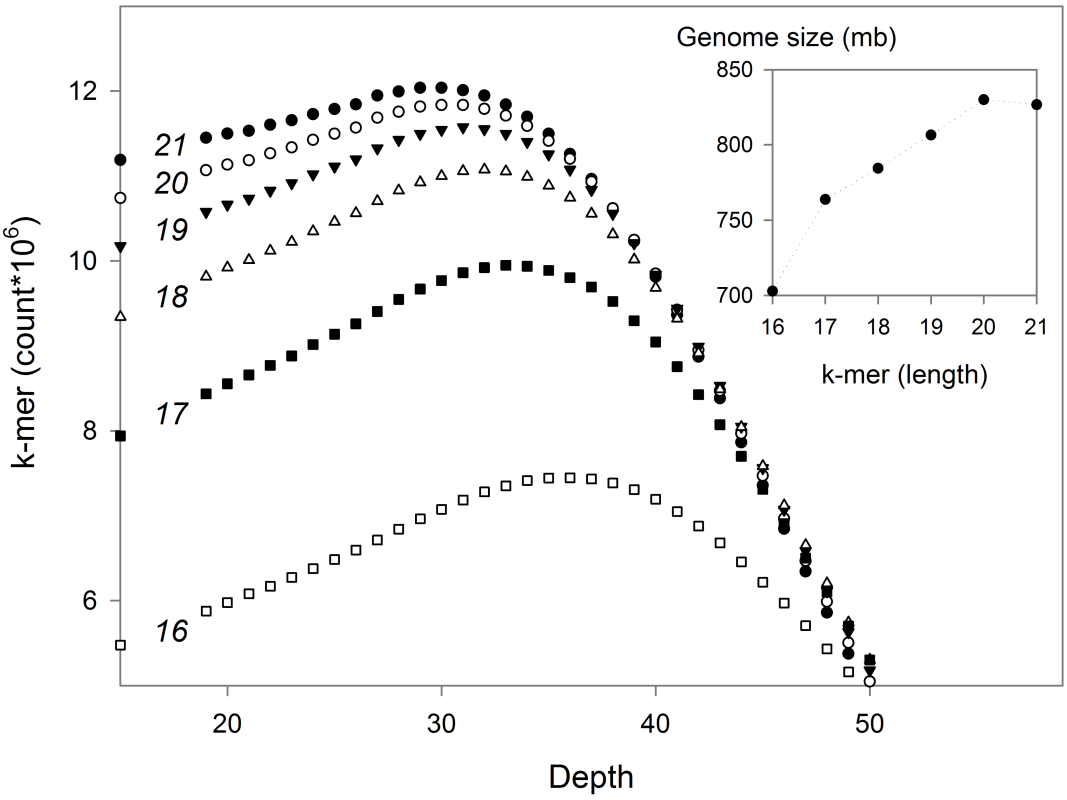
We obtained full-length Atlantic cod TLR protein sequences from the Newbler and Celera assembly through a combination of BLAST and Exonerate using the TLR protein sequences from the above-mentioned fishes as query. Both assemblies contained a similar set of TLR families. For further analyses, we chose the sequences from the Celera assembly (Supplementary Table 17), which were aligned together with Ensembl predicted TLR genes from Atlantic cod and stickleback to obtain reading-frames and exon-intron boundaries. All Atlantic cod protein sequences were aligned with the set of teleost and human TLR sequences using Mega4.0 (Version 4.1, Tamura et al. 2007). The alignment was visually inspected, adjusted and trimmed to contain the conserved regions (which included the TIR domain and additional upstream sequence). The final alignment contained 400 amino acids and is available upon request.

JTT+I+G+F was suggested as the best model of protein evolution for the TLR sequences by ProtTest3 (Version 2.4, Abascal et al. 2005). The maximum likelihood (ML) analysis was performed using RAxML Pthreads-parallelized version (Version 7.2.6, Stamatakis et al. 2005) with 8 threads, rapid bootstrap analysis and search for the best-scoring ML tree simultaneously, 100 replicates and using the model PROTCATJTT. Posterior probability scores were calculated using bayesian estimation of phylogeny by running the parallelized version of MrBayes3.12 (Version 3.1.2, Huelsenbeck et al. 2001, Ronquist and Huelsenbeck 2003). The substitution model was set to Jones, variable substitution rates across sites were accounted for by gamma distribution and invariable sites. The MCMC chains were carried out for approximately 3 million generations and trees were sampled every 100 generations. Posterior probability from the consensus of the best 1598 trees is presented on the maximum likelihood tree (Figure 4, Supplementary Figure 22).

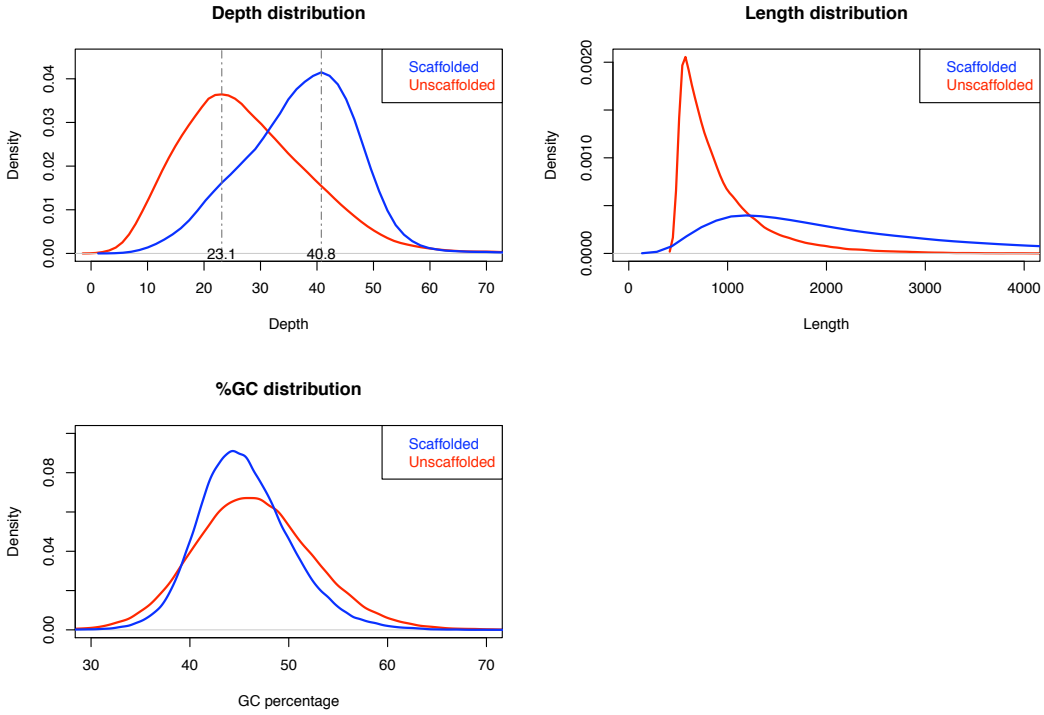
Supplementary Figures



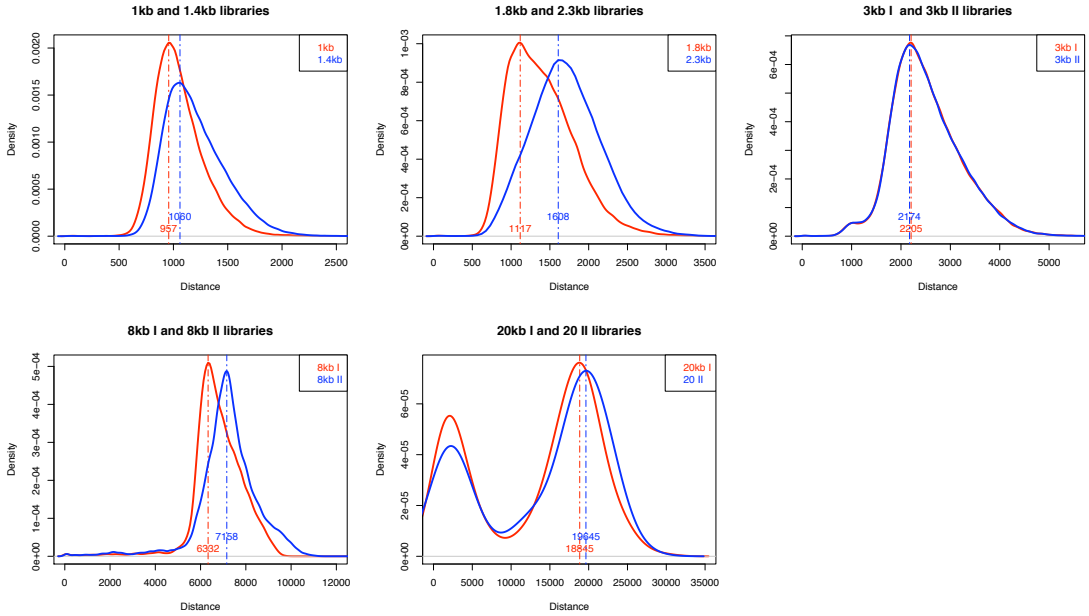
Supplementary Figure 1 Shotgun and paired-end read lengths.
Read length distribution for all shotgun (solid line) and paired end (stippled line) reads obtained from the NEAC_001. The peak read lengths were 503 bases for shotgun, and 389 bases for paired end reads, respectively.



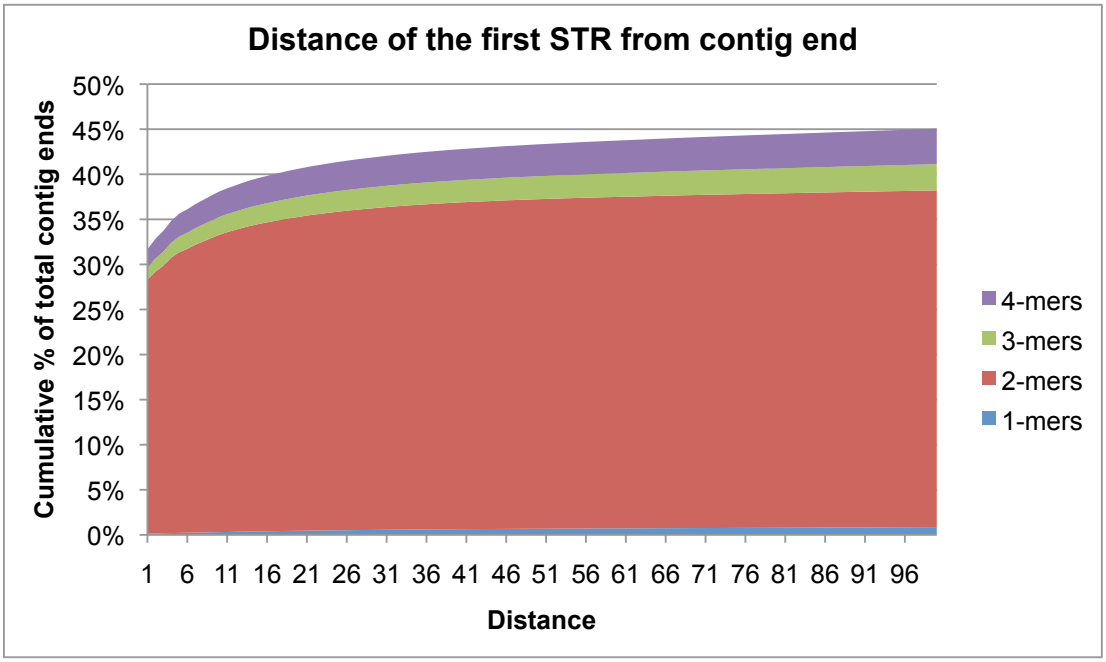
Supplementary Figure 2 K-mer based genome size estimation.
Distribution profiles of unique k-mer counts in the raw sequencing reads. Numbers in the graph indicate the different k-mer sizes in bases. The inset shows the genome size estimates based on the peak in the distribution profile for the k-mers of different size (see Supplementary Note 4 for formula).



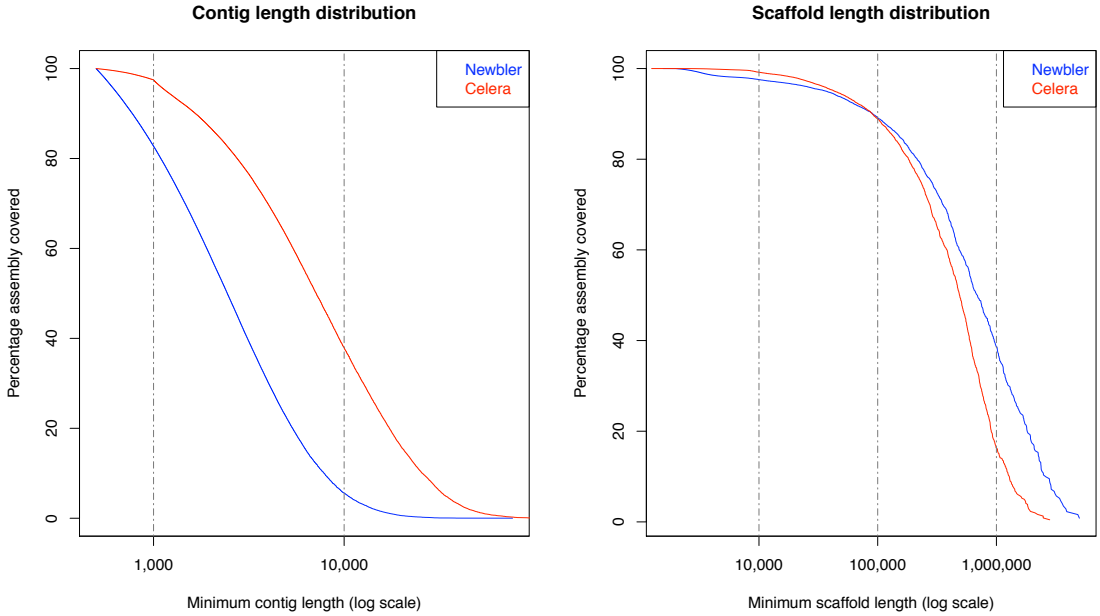
Supplementary Figure 3 Contig characteristics.
Density plots showing the per-contig read depth, contig length and per-contig percentage GC for scaffolded and unscaffolded contigs. The unscaffolded contigs show a lower read depth and are smaller. Their %GC distribution is comparable to that of the contigs in scaffolds.



Supplementary Figure 4 Density distributions of various paired-end libraries. Distances were calculated by mapping the paired end reads to the repeat-masked contigs using Newbler/gsMapper (settings: minimum overlap length 80% of the read length, minimum 96% identity in the overlap).

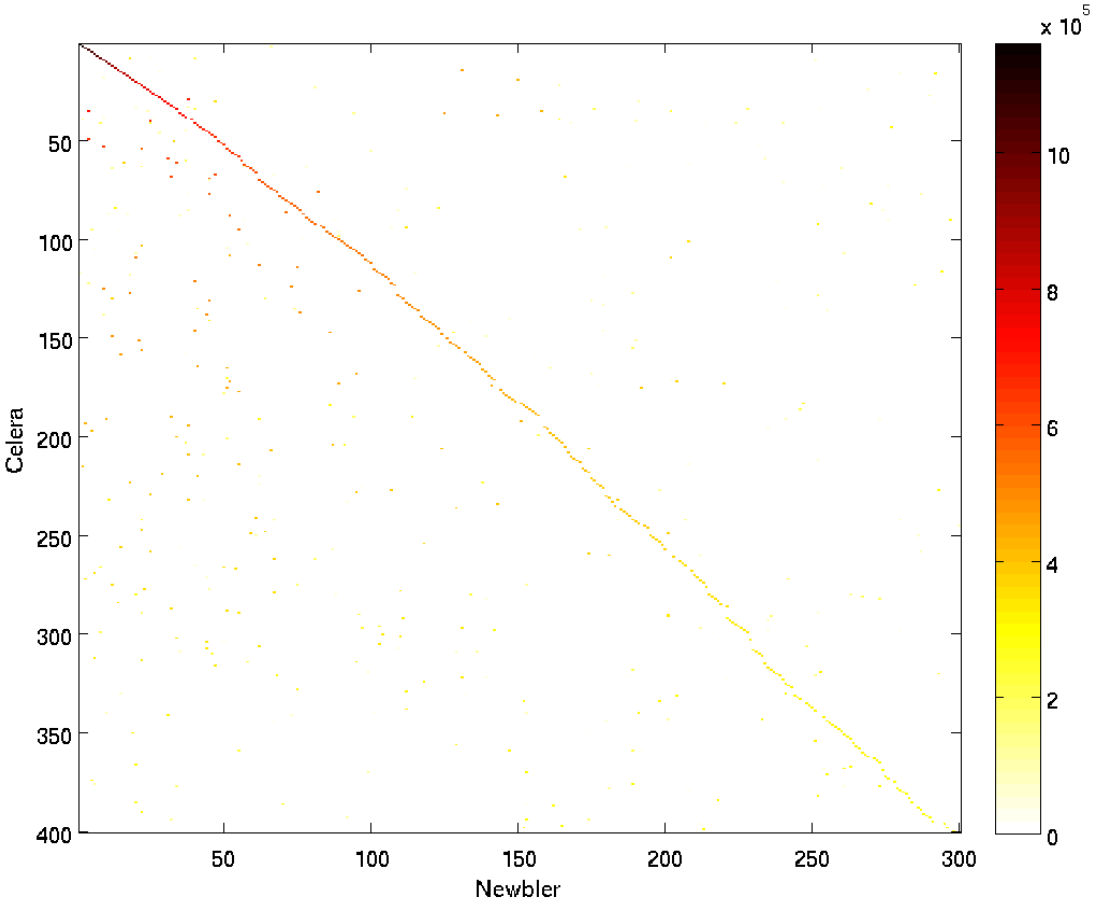


Supplementary Figure 5 Distribution short tandem repeats.
The graph shows, for the distance from the ends of contigs, the percentage of contig ends for which a short tandem repeat (STR, microsatellite) has started at or before that distance. 1-mers: homopolymers; 2-4 mers: di-, tri- and tetranucleotide repeats. More than 30% of contig ends have an STR starting at the first base.



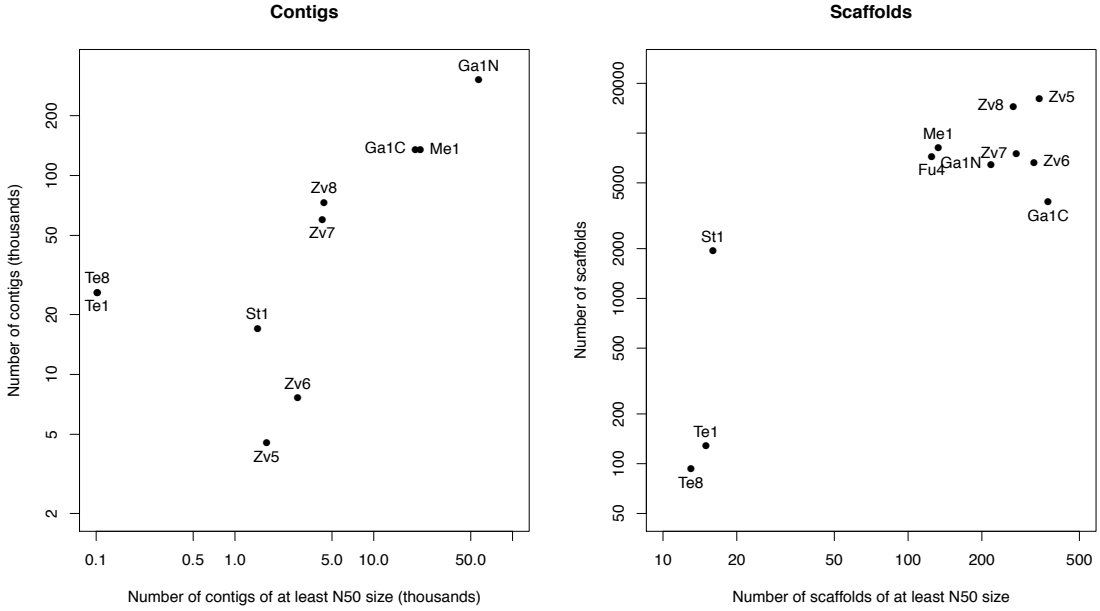
Supplementary Figure 6 Scaffold and contig size distributions of Newbler and Celera.

The percentage of the assembly included (y-axis) in contigs or scaffolds of a minimum size (x-axis, log scale) is shown for the Newbler (blue) and Celera (red) assemblies.

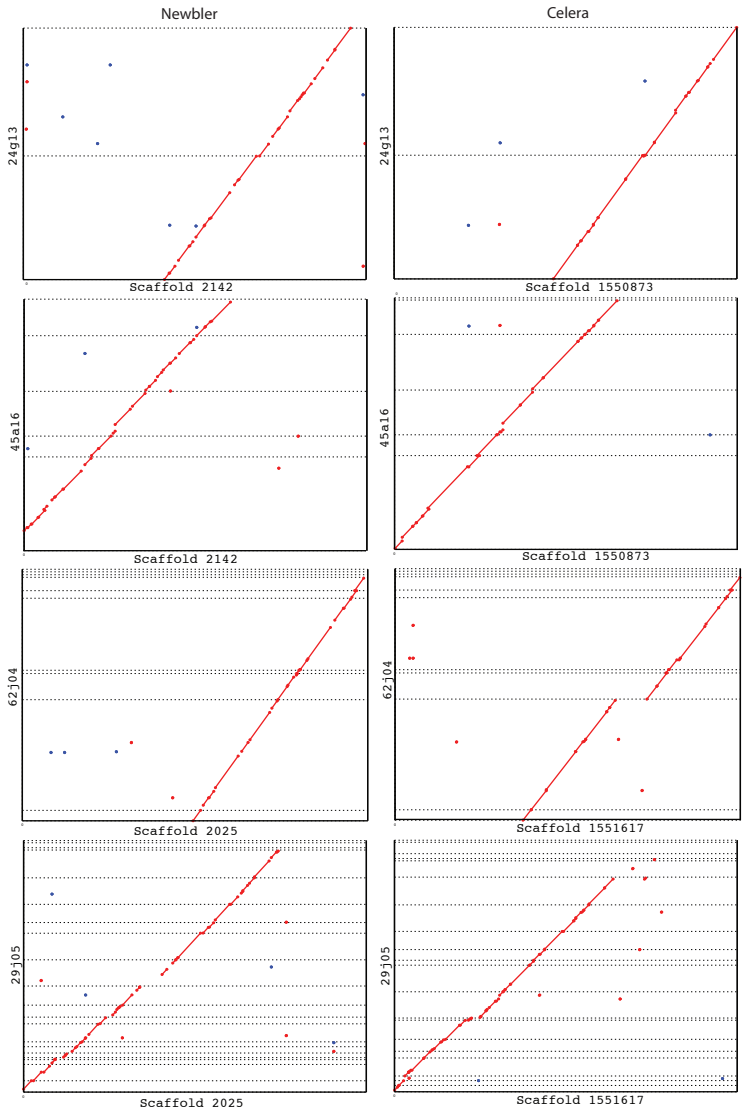


Supplementary Figure 7 Cumulated length matches between Newbler and Celera.

Density plot of the top 300 Newbler scaffolds (x-axis) and the top 400 Celera scaffolds (y-axis), both sorted in descending order. The colour-key represents the overlap length in basepairs.

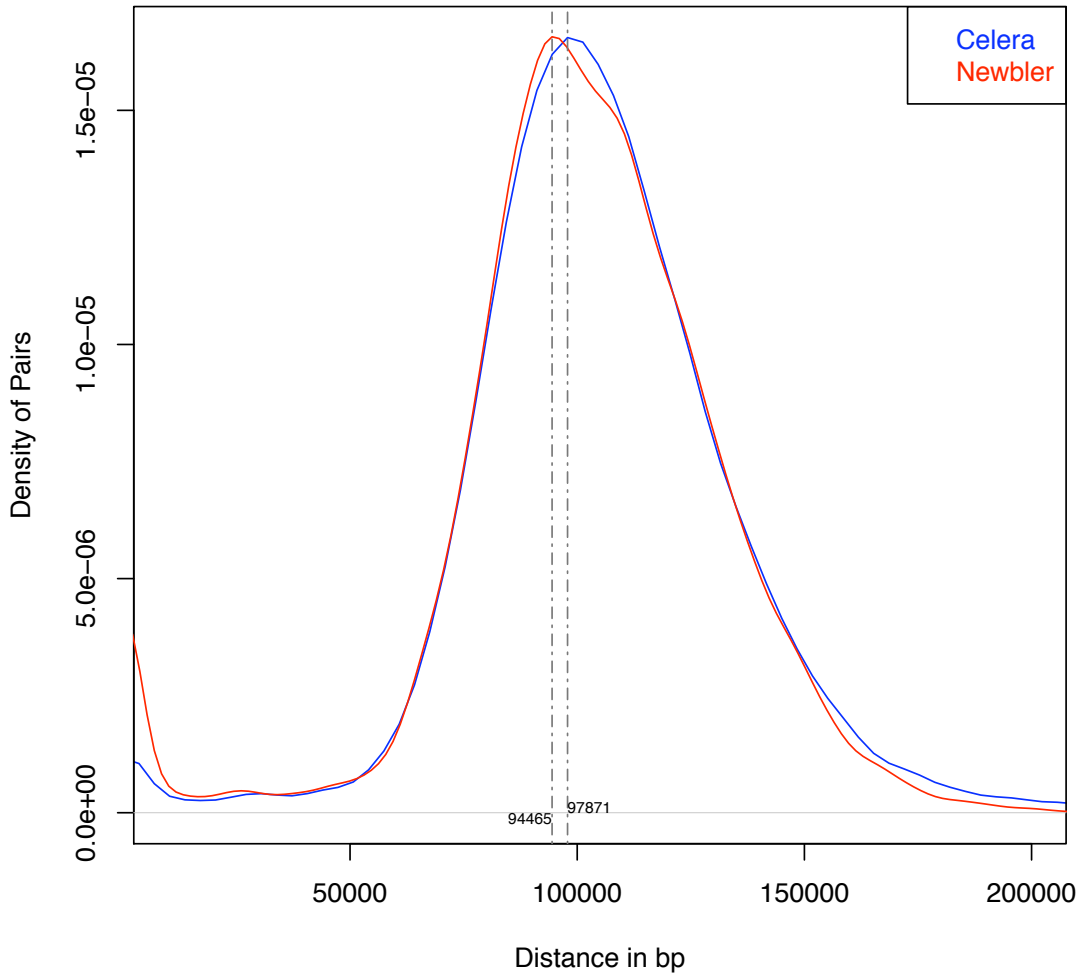


Supplementary Figure 8 Scaffold and contig metrics of teleost genomes.
Contigs and scaffolds metrics were based on the Ensembl databases for the Atlantic cod Newbler (Ga1N) and Celera (Ga1C) assemblies, and the following assemblies: *Tetraodon nigroviridis* versions 1 and 8 (Te1, Te8, the only versions present in Ensembl), *Oryzias latipes* (medaka) version 1 (Me1), *Danio rerio* (zebrafish) 5-8 (Zv5, Zv6, Zv7, Zv8), *Gasterosteus aculeatus* (stickleback) version 1 (St1) and *Takifugu rubripes* (fugu) version 4 (Fu4, no contig information was available for this assembly). Contig (left panel) and scaffold counts (y-axis, log scale) were plotted against the N50 number (number of contigs or scaffolds of at least N50 size, x-axis, log scale). For both metrics, lower numbers generally represent a more optimal assembly.

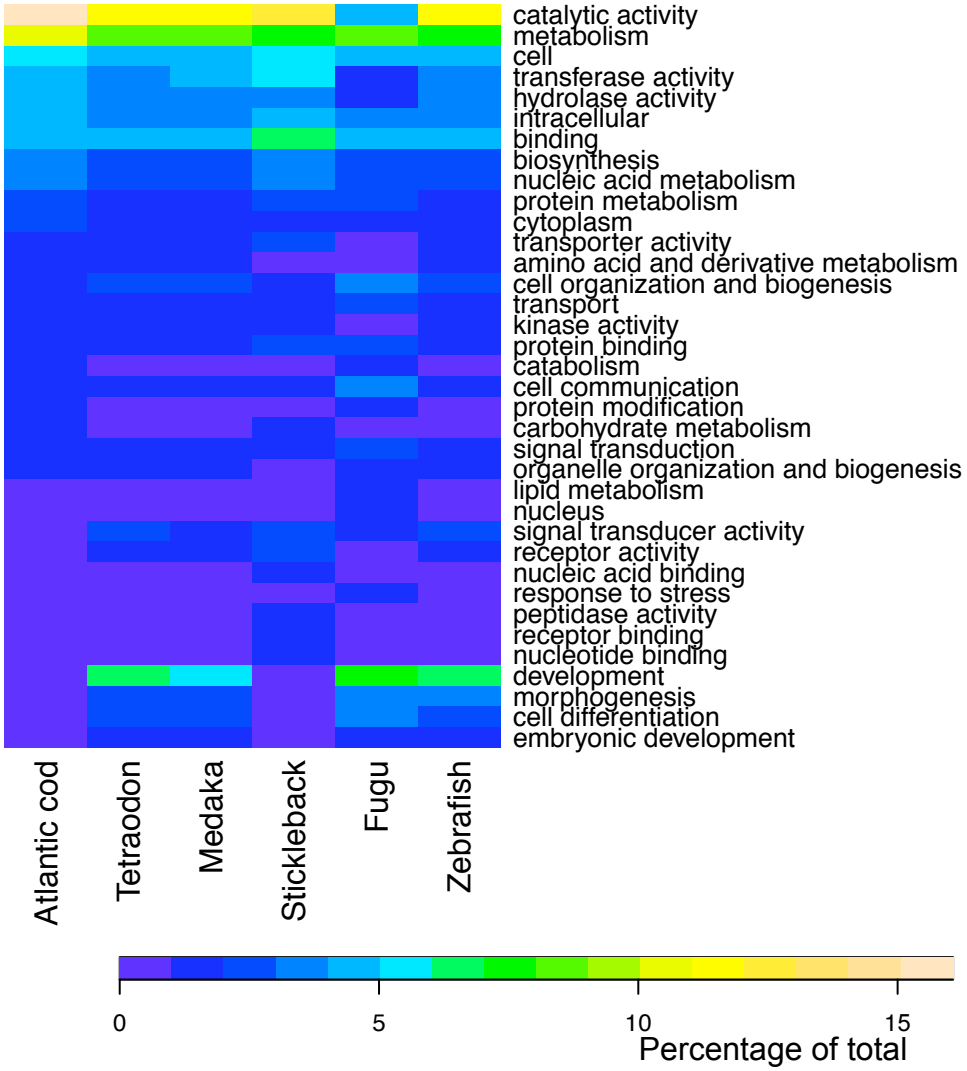


Supplementary Figure 9 BACs sequenced with Sanger

NUCmer alignments of the Newbler and Celera 454-genome assembly scaffolds (x-axis) and Sanger sequenced BAC scaffolds (y-axis) ordered and oriented such that the largest hits cluster near the main diagonal. Only alignments which represent the "best" one-to-one mapping are shown. Wherever the two sequences align, a coloured line or dot is plotted with forward matches in red, and reverse matches in blue. The dashed horizontal lines represent the borders between the BAC scaffolds. Large splits in the diagonal red line represent possible misassemblies (BAC 29j5 vs. Newbler scaffold 2025 and BAC 62j4 vs. Celera scaffold 1551617).



Supplementary Figure 10 Insert length distribution of BAC-ends.
Distances were calculated by mapping the repeat-masked BAC end sequences to the repeat-masked scaffolds of the Newbler and Celera assemblies using BLAST (red line, BLASTn, settings: maximum e-value 10^{-10}). The peak of the distribution is centred around the expected distance of 100kb between the paired ends.

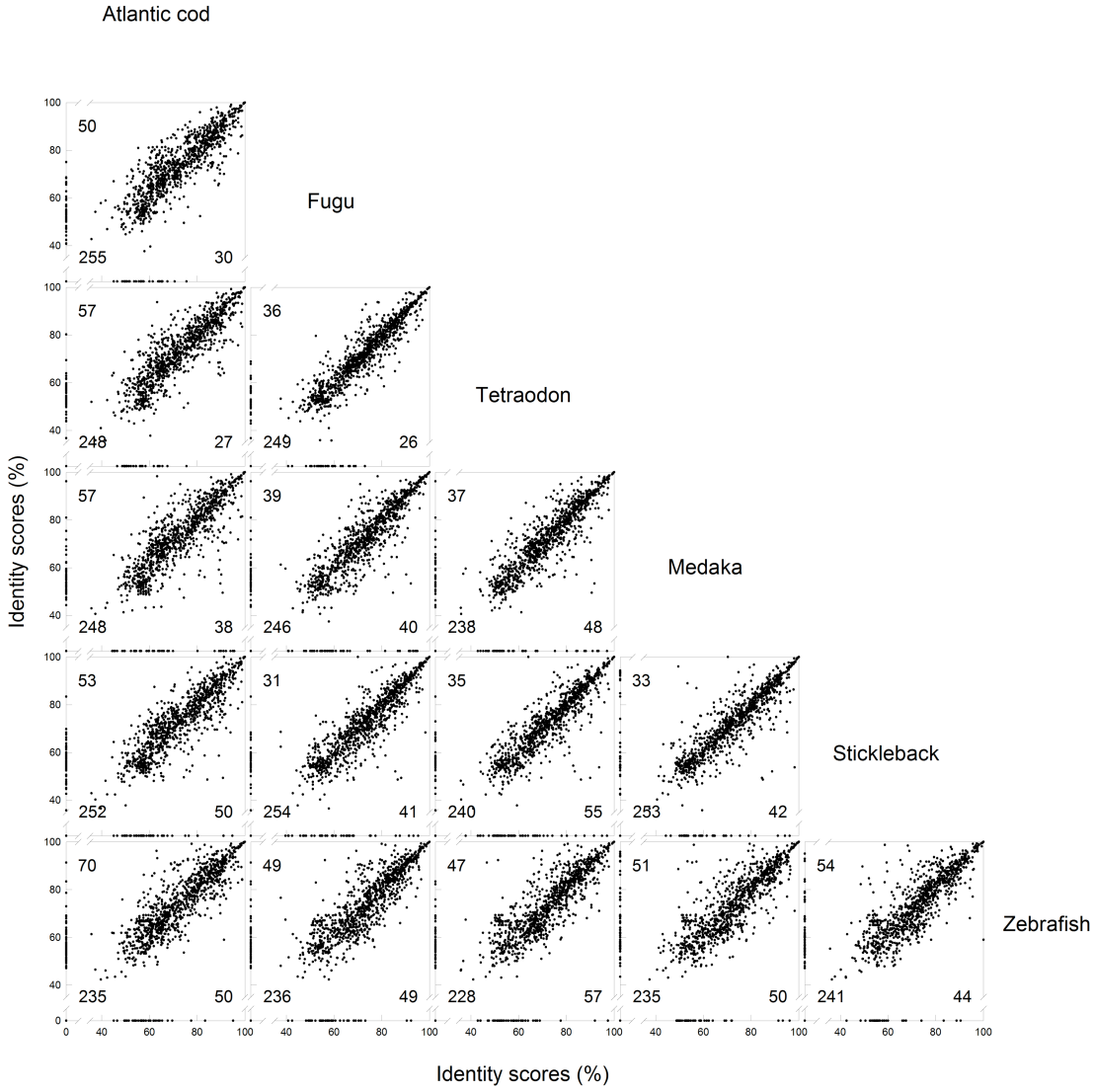


Supplementary Figure 11 Heatmap of genes associated with GO classes in teleosts.

Classes that have a percentage of 1% or higher in at least one of the teleost genomes are shown, sorted in descending order of the percentage in Atlantic cod.

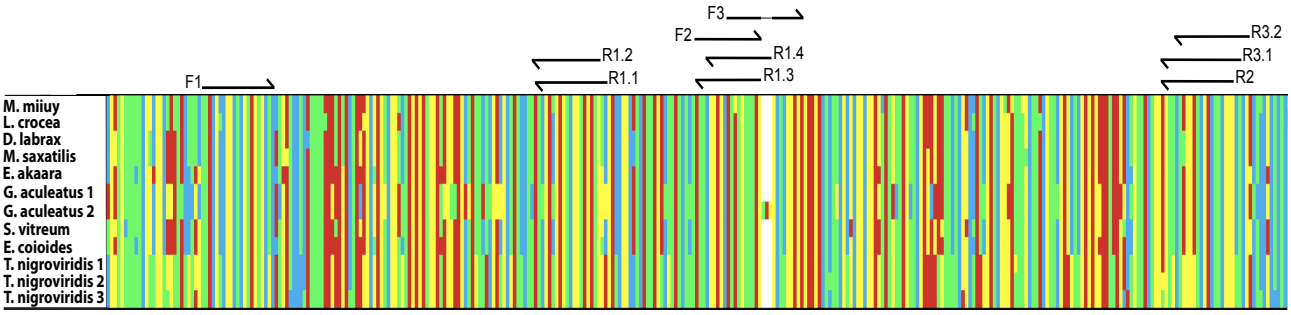
Supplementary Figure 12 Sample locations for eight Atlantic cod populations.

Tissue samples (muscle) were obtained in 2003/2004 and preserved in ethanol. Sample size is given in brackets.

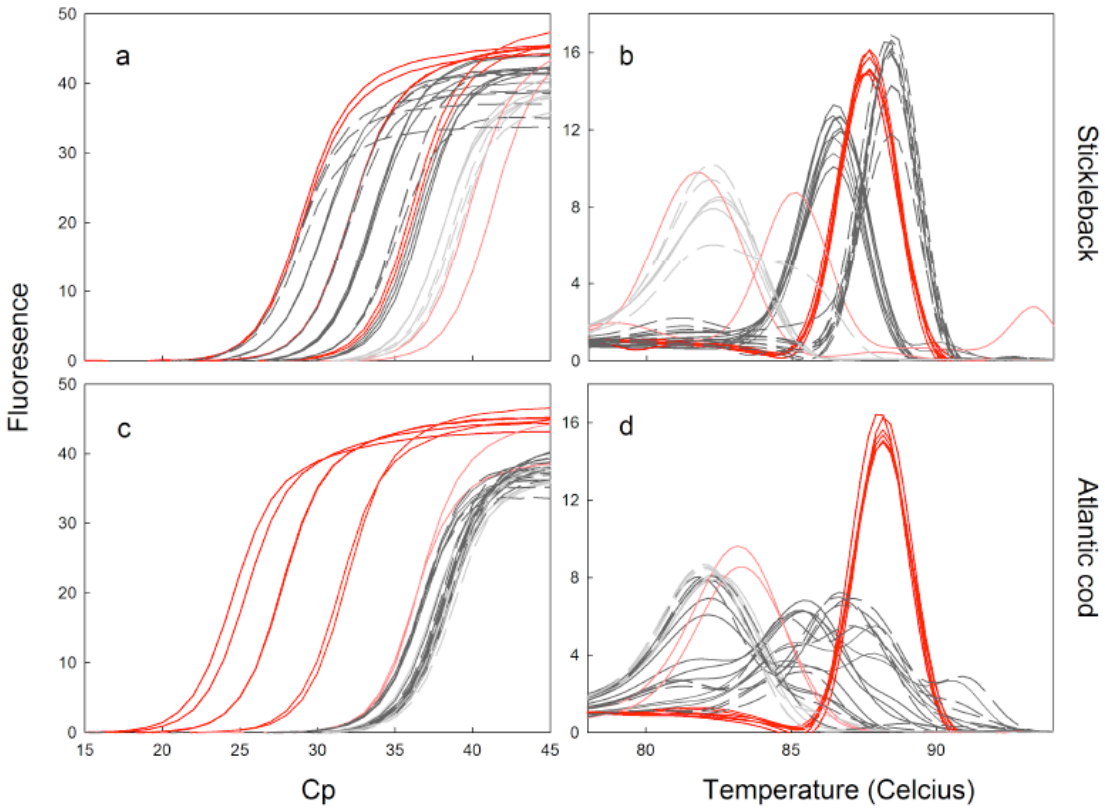


Supplementary Figure 13 Pairwise comparison of 1308 human immune genes.

Identity scores were compared for genes associated with GO:002376 (immune system process). Protein sequences were aligned to the teleost genomes using exonerate. The number of genes with no score in either teleost is shown in the bottom left corner of each panel. The number of genes with a score in one of the teleosts is plotted near its respective axis.

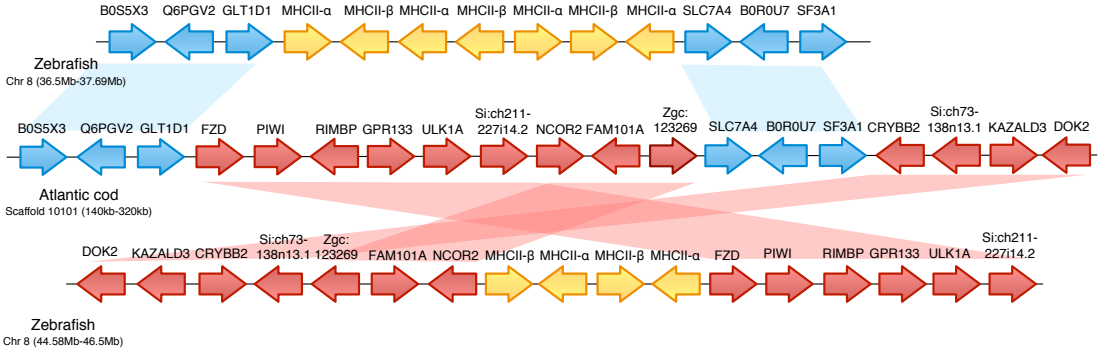


Supplementary Figure 14 Multiple alignment of MHCII Beta exon 3.
Several primers were designed to amplify parts of the MHCII B exon 3 using qPCR based on the sequences of selected teleosts (left column). Primer locations (arrows above) were recommended by Primer3 as implemented in NCBI and manually adjusted where necessary. Colors indicate the nucleotide bases guanine (yellow), cytosine (green), thymine (blue), adenine (red) and sequence gaps (white).



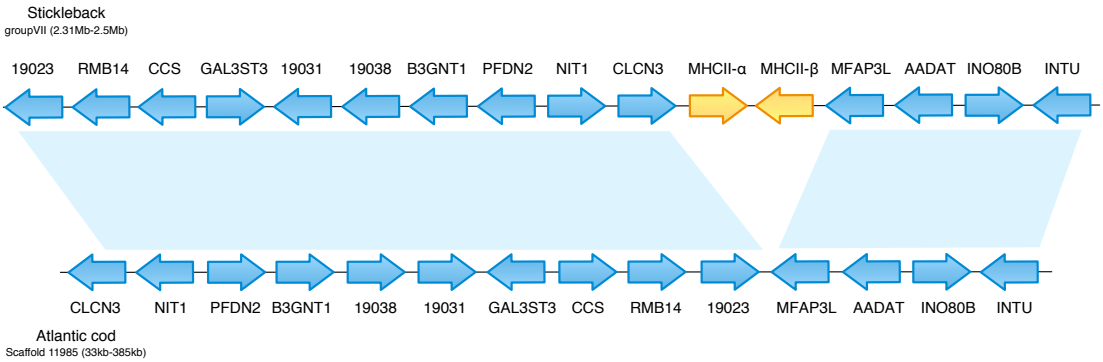
Supplementary Figure 15 qPCR amplification and melting point curves of MHCII and B2m.

Two MHCII primer pairs, Class_IIB_ex3_Forw1 + Class_IIB_ex3_Rev1.1 (grey) and Class_IIB_ex3_Forw1 + Class_IIB_ex3_Rev1.3 (dashed grey) were used to amplify Atlantic cod and stickleback. Positive controls using species-specific primers targeting Beta-2-microglobulin (B2m) were included (red). Results in the linear range (10X, 100X, 1000X) of the qPCR are shown. The MHCII primer-pairs were replicated three times for all dilutions (grey and dashed grey) and negative control (no template, light grey and dashed light grey), the B2m primers were replicated twice for all dilutions (red) and negative control (no template, light red). Cp refers to number of cycles.



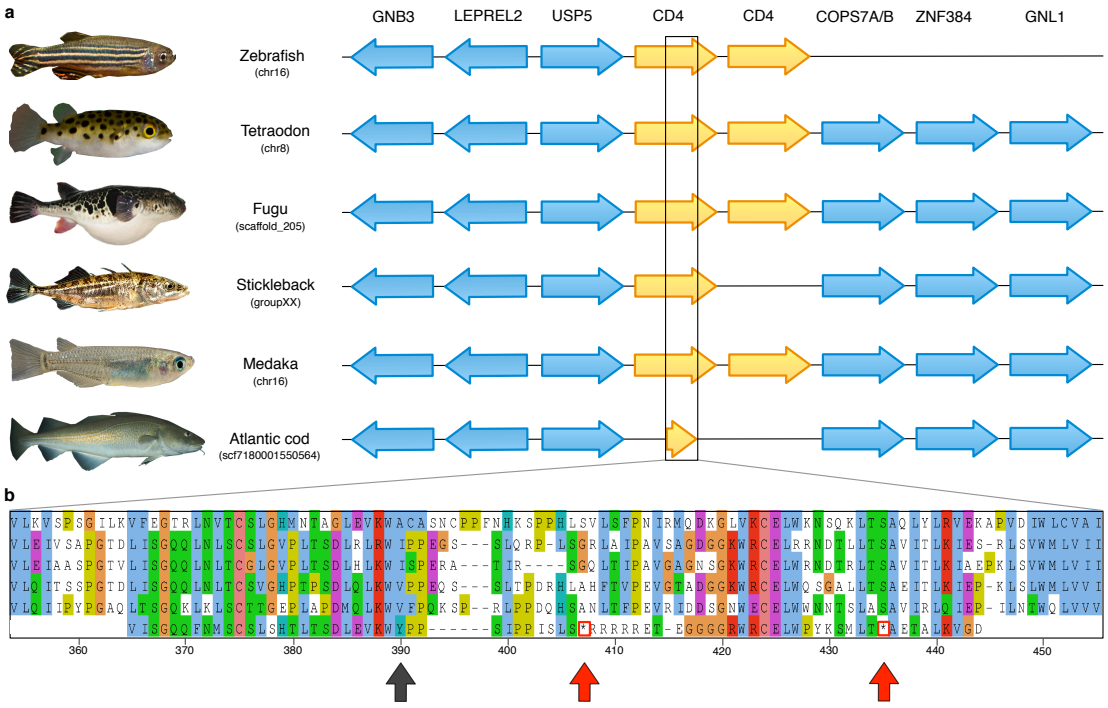
Supplementary Figure 16 Comparative alignments of zebrafish and Atlantic cod.

The gene order and transcriptional direction of flanking genes of the two MHCII regions (yellow arrows), located up- (blue arrows) and downstream (red arrows) from each other on chromosome 8 in zebrafish show evidence for multiple structural rearrangements in Atlantic cod. Both regions in zebrafish are drawn in the same transcriptional direction relative to each other. The locations of genes are plotted schematically. The flanking genes upstream and downstream of both MHCII regions in zebrafish occur in an aggregated set of syntenic blocks in a single region in Atlantic cod. The flanking regions of the upstream MHCII region in zebrafish show conserved order and transcription direction of the genes. Some genes in flanking regions of the downstream MHCII region in zebrafish have reversed order or transcriptional direction in Atlantic cod.



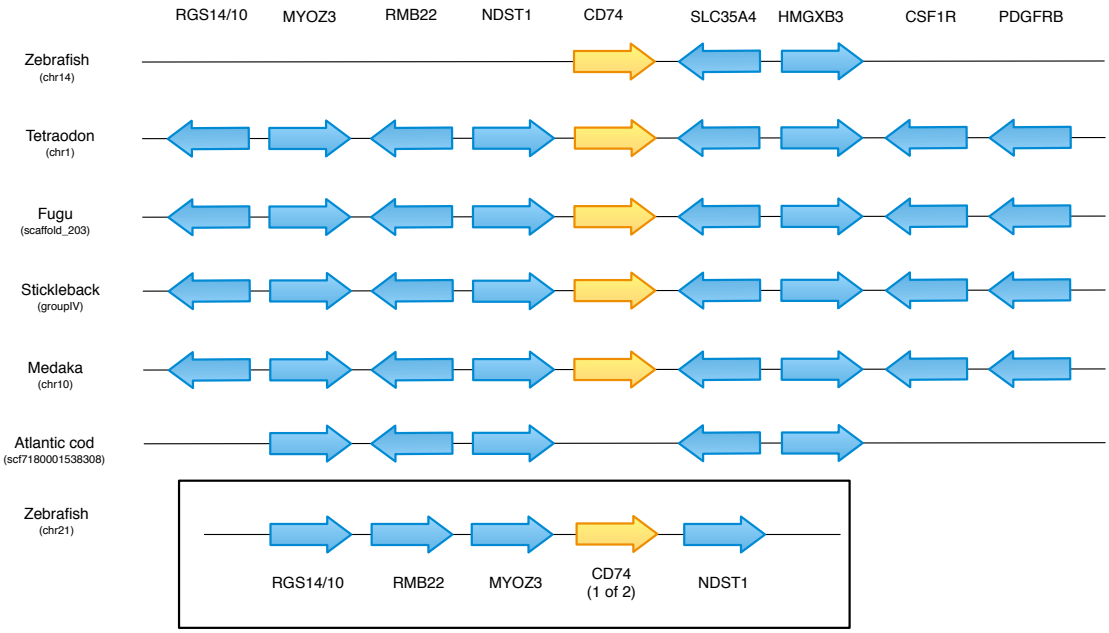
Supplementary Figure 17 Comparative alignment of stickleback and Atlantic cod.

The gene order and transcriptional direction of flanking genes (blue arrows) of the MHCII region in stickleback (yellow arrows) show evidence for structural rearrangement in Atlantic cod. The locations of genes are plotted schematically. The flanking genes upstream of the MHCII region in stickleback occur in reversed order in Atlantic cod, relative to the genes downstream. Within this reversed set of genes, order and transcription direction is conserved.



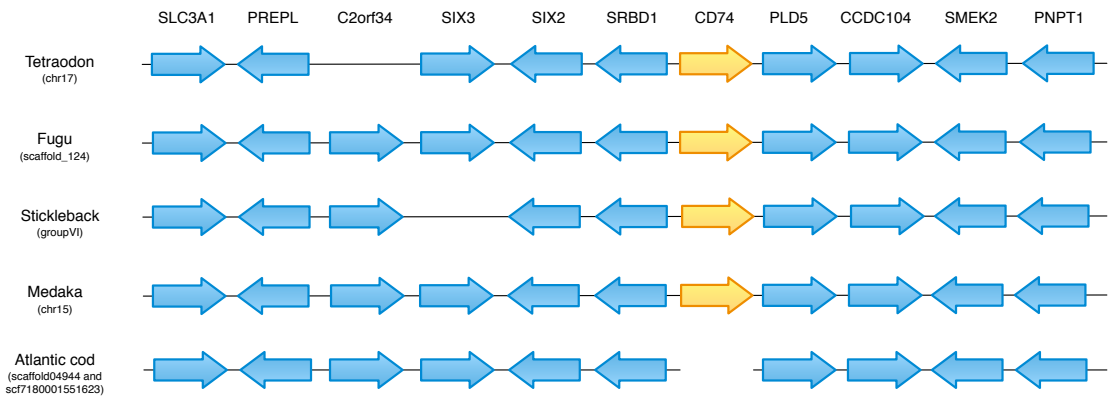
Supplementary Figure 18 Comparative alignments of CD4 region among teleosts

(**a**) The region around the CD4 genes (yellow arrows) shows conserved synteny of gene order and transcriptional direction of flanking genes (blue arrows). The locations of genes are plotted schematically. The Atlantic cod genome contains a fragment of CD4 (79 out of ~463 amino acids). Note that tetraodon has an additional set of genes for USP5 and CD4 upstream of COPS7A/B (not shown here). (**b**) Detail of the multi-species protein alignment showing the predicted protein sequence of the CD4 fragment. The translated protein sequence contains a frameshift (black arrow) and two stop codons (red arrows) in Atlantic cod. Chemical properties of the amino acids are shown using the standard ClustalX colour scheme.



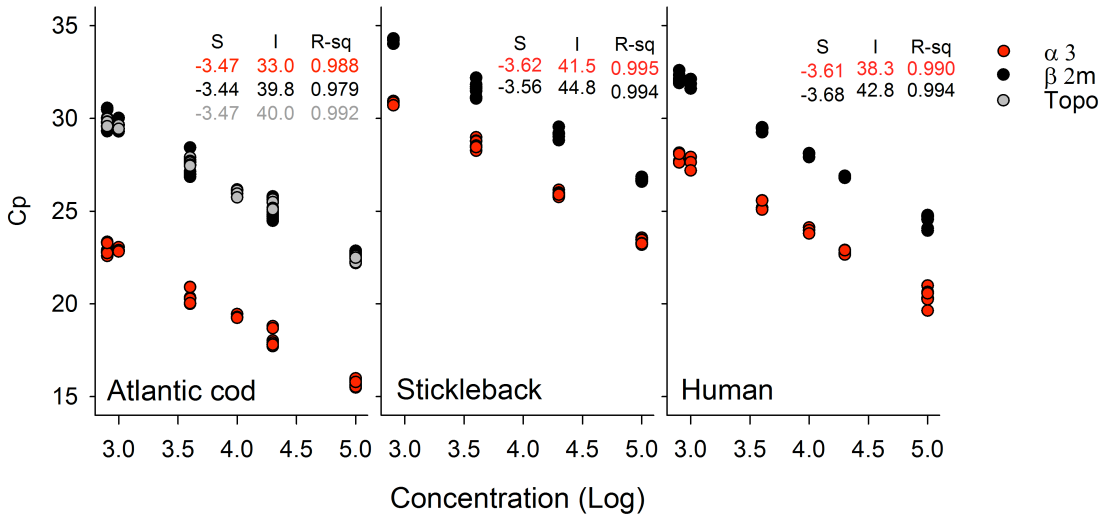
Supplementary Figure 19 Comparative alignments of CD74 region among teleosts.

The region around the invariant chain gene (CD74, yellow arrows) shows conserved synteny of gene order and transcriptional direction of flanking genes (blue arrows). The locations of genes are plotted schematically. The Atlantic cod genome lacks evidence for CD74. In zebrafish, homologs of CD74 are associated with either the upstream (framed inset, CD74) or the downstream flanking region.



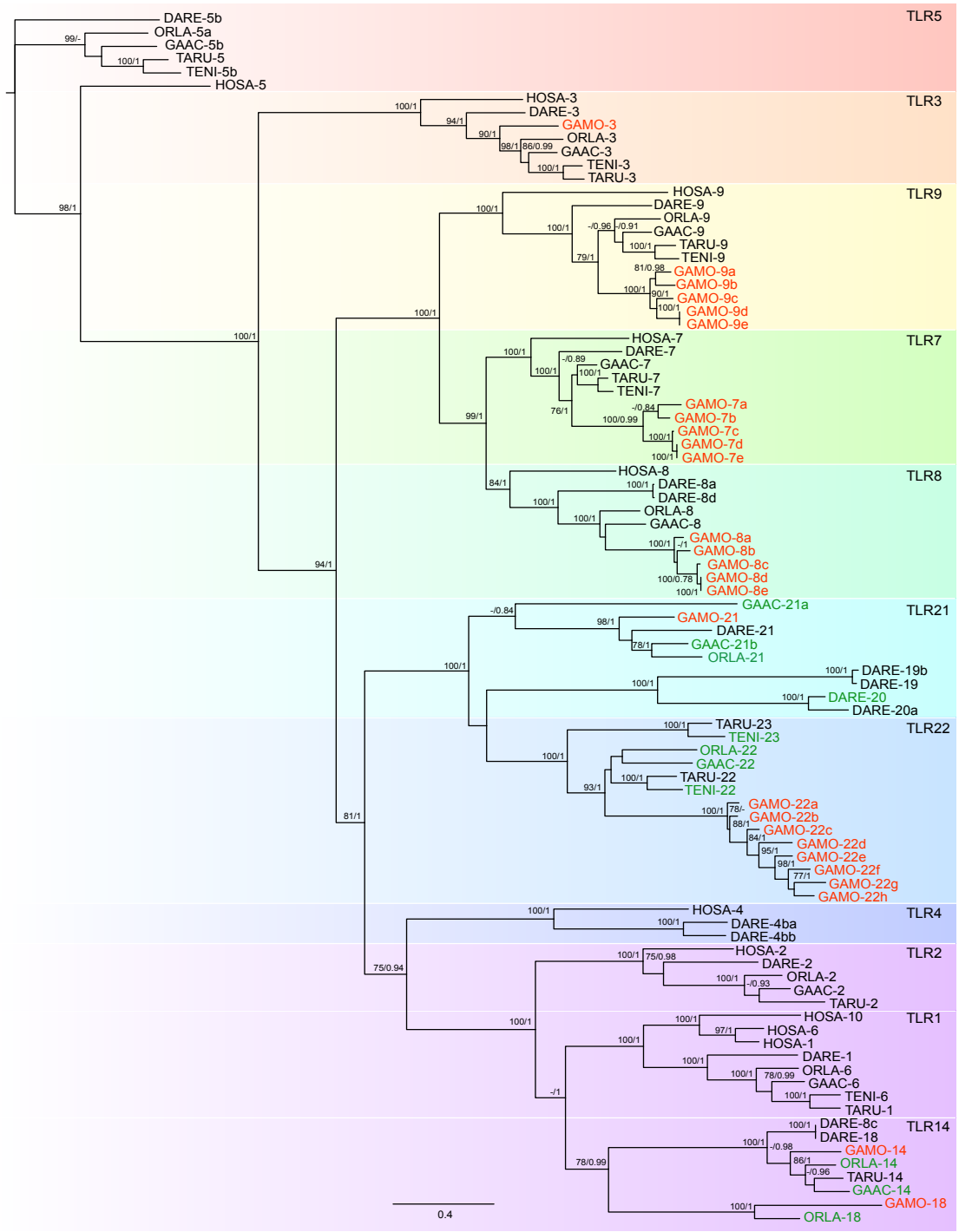
Supplementary Figure 20 Comparative alignments of CD74 region among tetraodon, fugu, stickleback, medaka and Atlantic cod.

The region around the invariant chain gene (CD74, yellow arrows) shows conserved synteny of gene order and transcriptional direction of flanking genes (blue arrows). The locations of genes are plotted schematically. The flanking regions in the Atlantic cod genome assemblies are resolved in two unconnected scaffolds. Both of these scaffolds lack evidence for a pseudogenic CD74.



Supplementary Figure 21 qPCR Cp-values for dilution series of genomic DNA samples of Atlantic cod, stickleback and human.

Species-specific primers were designed for the alpha3 sub-region of MHCI (α3), β-2-microglobulin (β2m) and topoisomerase (Topo). β2m and Topo represent single copy genes and α3 represents the multi-copy MCH I gene family in all three species. Slope (S), intercept (I) and fit (R-sq) are indicated for all respective regression lines.



Supplementary Figure 22 Phylogeny of Toll-like receptor (TLR) families.
TLR protein sequences were selected based on the conserved TIR domain for Atlantic cod (GAMO, red), including stickleback (GAAC), zebrafish (DARE), tetraodon (TENI), fugu (TARU), medaka (ORLA) and human (HOSA) as reference. Putative teleost TLR protein sequences (green) without specific TLR gene name in the Ensembl database were also included. Clades are assigned family name according to human orthologs where possible, or according to teleost orthologs nearest to Atlantic cod. Alignments were visually inspected and corrected where necessary. Maximum Likelihood (ML) values and bayesian posterior probabilities over 75/0.75 support the ML topology. Distance represents the average number of substitutions per site (scale).

Supplementary Tables

Supplementary Table 1 Number of sequenced 454 reads for different sequencing library types

Type of library	Libraries (n)	Shotgun reads (n)	Paired reads (n)	Pair distance (bp)[1]	Genome 'clone' coverage[2]
Shotgun	4	46,741,253			
1kb paired end	1	1,948,496	1,695,171	1,090	2.2
1.4kb paired end	1	1,760,881	1,357,361	1,248	2.0
1.8kb paired end	1	1,804,650	1,708,934	1,463	3.0
2.3kb paired end	1	3,281,715	2,235,804	1,791	4.8
3kb paired end	2	2,112,741	3,581,380	2,680	11.6
		1,750,435	2,709,567	2,676	8.7
8kb paired end	2	1,507,483	2,576,948	7,146	22.2
		1,754,508	3,162,796	7,686	29.3
20kb paired end	2	684,708	905,854	21,052	23.0
		234,185	241,774	21,081	6.1
Total	14	63,581,055	20,175,589		

[1]Distance reported by the Newbler program after assembly.
[2]Clone coverage is defined as the number of reads times the average paired end distance divided by the genome size. It represents the coverage when the pair distance is taken as the read length.

Supplementary Table 2 454 reads and Sanger EST reads from cDNA of several *Gadus morhua* tissue types used for the transcriptome assembly.

	NEAC_001[2]		Other *G. morhua*[3]	Total[1]
Tissue	**454**	**454**	**Sanger[4]**	
Brain	257,456		13,811	271,267
Eggs	141,381		11,360	152,741
Gonad	117,701		14,573	132,274
Headkidney			18,569	18,569
Spleen	171,144		3,473	174,617
Heart			16,442	16,442
Headkidney +heart	43,759			43,759
Headkidney + spleen			5,366	5,366
Hindgut	213,694			213,694
Liver	425,640		23,246	448,886
Pylorus			2,557	2,557
Pyloric caecae or anterior stomach			5,855	5,855
Intestine			7,391	7,391
Bone and muscle		52,806[5]	4,013	56,819
Pituitary gland			1,682	1,682
Gill			7,272	7,272
Virus infected		55,643		55,643
Several developmental stages[6]		214,130		
Unknown			27,716	27,716
Total	1,370,775	108,449	163,326	1,642,550

[1]Reads were filtered by trimming low quality bases (average Phred score less than 20 in a 20 bp window) and discarding reads shorter than 100 bp.
[2]Represents the sequenced specimen.
[3]Represents other *G. morhua* individuals.
[4]Made available through the Cod Genome Consortium. The data set originated from studies carried out at the Institute of Marine Research, Norway (Edvardsen et al 2010), NIFES (National Institute of Nutrition and Seafood Research) (Lie et al. 2009, Olsvik and Holen 2009) and groups affiliated to the Cod Genomics and Broodstock Development, Canada (unpublished data, http://codgene.ca).
[5]cDNA reads from the bone- and muscle tissue of cod larvae (Mari Moren et al, unpublished data). We selected for reads with high sequence identity (98% sequence identity over 75% of the read) compared to the cod reference sequence to avoid incorporating erroneous reads from contamination.
[6]From (Johansen et al. 2009).

Supplementary Table 3 Summary of transcriptome assembly statistics

Metric	Count
Number of assembled reads	793,328
Number of singleton reads	248,121
Number of reads classified as repeats	162,086
Number of contigs	46,400
Number of contigs, quality trimmed	41,419
Number of bases, quality trimmed contigs	28Mb

Supplementary Table 4 Statistics from the alignment of the transcriptome assembly to the genome assemblies

Genome assembly	Newbler assembly contigs	Newbler assembly scaffolds	Celera assembly
Transcriptome contigs aligned[1]	41,223 (99.5%)	38,965 (94%)	40,599 (98.0%)
Median alignment coverage[2]	99.8%	98.6%	99.6%
Directional alignment errors[3]	ND[5]	1,324 (3.2%)	1,623 (3.9%)
Positional alignment errors[4]	ND[5]	646 (1.6%)	609 (1.5%)

[1]BLASTn maximum e-value 10^{-9}

[2]By adding the lengths of the high-scoring segment pairs (HSPs) below the maximum expectation value from the transcriptome-to-genome alignments, the total length of aligned sequence in each transcriptome contig was calculated. This length divided by the total length of each contig was defined as its "alignment coverage".

[3]Alignment contains HSPs oriented in opposite directions relative to each other, below the maximum expectation value.

[4]Alignment contains HSPs in consistent directions but with internally inconsistent positional mapping.

[5]ND, Not Determined.

Supplementary Table 5 Sanger-sequenced BAC assemblies

BAC	24g13	29j05	45a16	62j04
Reads[1]	1,658	1,796	3,273	1,500
Bases[1]	1,587,959	1,965,392	3,352,374	1,482,045
Assembly coverage[2]	10.8x	7.4x	16.4x	8.4x
Total assembly length (bp)	108,611	129,499	143,697	100,791
Scaffold size classes				
>50kb	2		1	
10-50kb		4	4	2
1-10kb		16	1	7

[1]Represents the number of unassembled, unfiltered reads and corresponding number of bases that were used as input for the assembly.
[2]As reported by the Celera assembly program.

Supplementary Table 6 TE elements in the Atlantic cod genome

TE element[1]	Teleost (Repbase)	*de novo*	Teleost (Repbase) & *de novo*
SINE	0.25	0.55	0.58
LINE	1.21	2.94	3.3
LTR	1.45	3.90	4.88
DNA	1.81	5.21	6.39
Unclassified	0.01	2.81	2.81
Small RNA	0.04	0.04	0.07
Satellites	0.03	0	0.03
Simple repeats	5.92	6.63	5.92
Low complexity	1.59	1.57	1.59
Total	12.31	23.49	25.40

[1]The abundance of TE elements (%) in the assembled genome (contigs longer than 500 bp) were identified using the teleost Repbase, a *de novo* repeat library, or a combination of both and masked using RepeatMasker.

Supplementary Table 7 Annotation Statistics

Annotation	Number	Bases (Mb)	Assembly (%)
Protein coding genes	20,095	28.1[a]	5.27
Pseudogenes	518	0.38	0.07
rRNA	590	0.06	0.01
miRNA	414	0.03	0.01
snoRNA	382	0.05	0.01
snRNA	115	0.01	0.003
miscRNA	40	0.01	0.002
Transposable elements	614,494	137	19.74%
Simple and low complexity	731,280	57	8.21%

[a]excluding introns

Supplementary Table 8 Comparison of assemblies mapped to SNP linkage map

Assembly	SNPs (n, %)		Scaffolds[1] (n)			Sequence Linked (Mb)
	All (1694[2])	Linkage(924[3])	Mapped[4]	Mapped 2+[5]	Consistent[6]	
Newbler	1609, *95.0*	879, *95.1*	455	178	175 (98.3%)	332
Celera	1604, *94.7*	881, *95.4*	536	200	149 (74.5%)	274

[1]Includes 145,150 contigs that were not in scaffolds.
[2]Total number of SNPs available.
[3]Number of SNPs for which linkage information is available.
[4]Scaffolds mapped to one or more SNPs.
[5]Scaffolds mapped to two or more SNPs.
[6]Scaffolds mapped to two or more SNPs associated with the same linkage group.

Supplementary Table 9 Counts of orthologous genes present on 23 linkage groups of Atlantic cod

[1]Stickleback chromosomes

	Atlantic cod linkage groups																						
Chr[1]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
XII	291	0	3	0	0	0	20	2	0	0	0	0	24	0	0	0	0	2	1	0	0	0	0
XI	0	267	10	0	0	1	2	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	2
IX	0	4	258	1	0	1	1	1	0	1	1	0	0	0	4	3	1	1	0	2	0	0	0
XIV	0	1	0	163	0	9	0	2	0	0	0	1	0	0	0	1	1	0	101	0	0	1	0
XV	0	0	0	0	227	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	8	0	0
XIII	2	0	0	8	0	261	1	0	0	0	22	2	0	1	1	0	1	1	7	0	0	0	0
VII	1	0	5	0	1	1	245	3	0	2	1	1	1	0	0	2	92	2	0	0	1	1	0
III	0	0	0	0	0	1	0	263	0	1	0	7	0	0	0	2	0	0	0	2	0	0	3
XIX	3	24	1	7	0	3	1	0	279	0	0	0	6	14	0	3	1	2	11	1	1	1	0
IV	5	1	2	10	2	0	3	1	4	186	1	1	2	0	1	0	0	0	48	2	0	0	1
XX	7	0	1	0	0	22	1	0	0	0	196	0	2	46	0	0	0	0	0	0	1	12	2
VIII	1	0	2	1	2	0	1	5	0	0	0	241	0	2	0	0	0	0	0	0	0	0	0
XVII	43	0	1	2	0	0	2	0	0	0	0	0	189	0	0	1	0	0	0	0	0	1	2
II	0	0	3	1	1	0	1	1	10	0	1	3	0	288	0	0	1	1	0	1	1	0	0
VI	0	0	3	2	0	0	2	1	0	0	3	3	0	0	173	0	2	2	0	1	1	0	1
I	0	1	2	7	3	0	19	1	0	0	1	2	1	1	1	222	3	0	1	16	0	1	0
V	0	7	0	0	1	0	5	1	0	0	0	0	0	0	3	0	1	197	0	0	0	0	0
XVI	1	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	4	1	0	187	1	0	1
XVIII	1	0	1	1	12	1	0	0	0	0	4	2	0	0	1	2	0	3	0	0	216	0	0
X	0	2	1	0	2	1	0	1	0	0	6	1	1	3	1	0	0	0	0	2	1	162	0
XXI	0	0	2	0	0	0	3	9	1	0	0	1	3	0	0	0	0	8	0	1	0	0	89

Supplementary Table 9 – continued
[2]Tetraodon chromosomes

	Atlantic cod linkage groups																						
Chr[2]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
9	230	0	13	1	1	0	3	3	2	1	1	0	11	1	0	1	0	0	2	0	0	2	0
3	4	161	7	2	1	1	10	9	0	1	7	2	2	0	0	0	5	17	0	16	1	1	1
18	0	28	180	3	0	3	2	2	1	3	1	0	0	1	4	1	2	0	6	0	0	0	3
4	0	1	2	93	0	18	2	2	1	2	2	0	0	1	2	1	0	1	94	0	0	1	2
10	0	0	2	0	192	0	3	1	0	2	0	0	0	0	3	23	2	1	0	0	16	1	1
12	1	1	2	9	0	181	1	1	0	0	9	1	2	1	1	2	0	2	2	0	1	1	0
7	0	3	3	2	0	1	176	3	2	3	3	1	1	0	1	10	7	2	0	2	0	0	0
15	0	2	1	3	0	0	0	112	1	0	0	16	0	2	0	5	4	9	0	1	0	3	5
13	1	2	0	1	0	1	0	1	205	0	1	0	1	17	0	0	0	1	3	0	0	0	0
1	1	2	5	5	0	1	3	24	1	142	2	189	0	2	0	2	2	1	0	1	0	0	0
8	7	0	1	0	0	14	0	0	3	0	129	1	1	34	0	0	0	0	1	0	0	18	1
11	39	0	0	1	0	0	0	3	0	0	2	0	206	1	0	0	0	1	1	1	0	3	0
5	1	0	2	4	0	2	0	1	9	1	2	1	1	209	0	1	1	1	0	4	1	0	1
17	0	1	7	1	1	1	0	1	0	0	2	1	0	0	131	2	1	13	0	0	1	1	0
16	0	0	0	0	0	1	10	5	0	1	1	1	0	0	1	97	1	1	0	0	0	0	0
20	3	0	4	0	0	0	1	0	0	0	0	0	0	0	2	1	11	0	0	0	0	0	0
2	1	11	3	1	1	1	2	4	1	0	2	0	0	2	2	0	7	130	0	174	2	3	2
14	1	0	0	1	11	0	1	2	0	0	1	1	0	2	1	1	0	2	0	0	178	1	1
21	0	0	2	0	3	0	0	1	3	0	5	0	0	2	0	0	0	0	0	0	1	90	0
6	1	0	0	0	0	0	2	12	2	0	0	0	1	0	0	0	0	2	0	0	0	0	67
19	3	2	0	14	0	1	2	0	4	0	1	1	1	1	0	0	0	1	43	0	0	0	0

Supplementary Table 9 – continued
[3]Medaka chromosomes

	Atlantic cod linkage groups																						
Chr[3]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
7	247	0	11	0	0	0	7	1	1	0	2	0	13	0	0	0	0	0	0	0	0	0	0
8	1	184	12	0	0	1	1	1	0	0	1	0	0	2	0	0	0	8	0	0	1	0	1
1	0	25	180	2	1	4	1	2	0	2	0	0	0	0	3	1	1	1	5	0	0	0	0
12	0	0	1	119	0	11	1	0	0	0	1	3	0	0	1	1	2	0	93	0	0	1	0
22	0	0	1	0	177	0	0	0	0	1	0	2	0	0	1	2	0	0	0	1	10	4	0
9	0	0	0	11	0	165	2	1	0	0	12	3	0	0	1	0	0	0	3	1	0	1	0
14	1	1	1	0	0	0	203	1	0	1	1	1	0	0	0	9	3	1	1	3	0	0	0
17	0	0	0	0	0	0	0	198	0	0	0	9	1	1	0	3	0	0	2	0	0	1	4
6	0	0	0	0	0	0	1	0	175	1	1	0	0	19	0	0	0	0	3	0	0	0	1
10	0	1	3	0	0	0	3	0	1	141	0	1	0	0	0	0	1	0	0	0	0	0	0
16	3	0	1	0	0	15	2	0	0	0	148	0	0	25	0	0	0	0	1	0	2	11	0
4	1	0	0	2	0	0	0	10	2	0	0	218	0	1	0	0	0	0	0	0	0	0	0
5	30	0	0	0	0	0	2	2	1	0	0	0	192	1	0	2	1	0	1	0	0	0	1
3	0	0	2	2	1	2	0	1	10	1	0	0	0	191	1	0	1	0	0	3	1	1	0
15	0	0	4	1	0	0	0	0	0	0	2	1	0	0	142	0	2	2	0	0	2	0	0
13	0	0	1	1	0	1	13	0	1	0	1	0	1	0	1	176	0	0	0	0	0	0	0
18	0	1	2	0	0	0	1	2	0	0	0	0	0	0	0	2	81	0	0	1	0	0	0
19	0	10	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	139	0	0	0	0	0
21	0	0	1	0	0	0	1	1	0	0	0	0	0	2	0	0	2	0	1	139	1	1	0
24	0	0	0	0	8	0	0	0	1	1	0	1	0	1	1	1	1	1	0	0	162	1	0
11	0	1	1	0	1	0	0	1	0	0	14	1	0	1	1	0	0	0	0	1	0	140	1
20	0	0	0	0	0	0	1	6	0	0	1	0	0	0	0	0	0	8	0	0	0	0	72
2	0	0	0	5	1	0	7	3	0	2	1	4	0	0	0	0	1	1	0	12	0	0	1
23	2	0	0	14	0	0	0	0	13	0	2	0	3	1	0	0	0	0	37	0	0	0	0

Supplementary Table 9 – continued
[4]Zebrafish chromosomes

	Atlantic cod linkage groups																						
Chr[4]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
23	98	0	10	0	0	3	8	3	1	1	0	5	8	0	1	2	1	0	0	0	3	1	1
3	1	149	15	0	0	2	1	1	0	0	2	0	0	2	1	2	2	25	0	0	0	1	1
1	0	7	76	0	0	1	0	4	0	6	0	1	1	6	3	1	2	2	0	4	0	1	0
10	0	0	3	35	1	17	33	1	1	2	1	1	2	2	4	10	2	1	29	2	0	0	1
17	2	1	2	0	62	0	0	1	0	0	2	0	0	8	19	0	0	5	0	0	17	3	0
8	60	1	1	8	0	66	4	1	0	0	7	24	11	0	1	1	0	1	7	1	0	0	0
21	0	0	1	22	1	9	51	0	2	7	2	0	4	0	0	9	5	0	24	0	0	0	0
2	1	3	0	1	6	2	2	96	1	0	1	15	1	0	0	3	0	0	0	0	0	2	9
25	0	11	0	2	0	2	1	1	81	0	0	0	0	24	0	2	0	0	4	0	0	0	0
14	0	0	6	1	4	1	10	0	1	83	0	0	0	1	1	2	0	1	0	2	0	0	0
16	1	0	0	1	1	7	0	1	2	0	90	1	0	21	2	1	0	0	0	2	3	10	1
22	5	0	10	3	1	0	1	5	1	0	0	39	7	0	1	0	0	0	2	7	1	1	0
11	28	1	7	1	2	0	1	5	0	0	1	18	53	5	0	0	0	2	1	2	0	1	0
7	6	7	1	3	1	3	2	3	13	1	1	0	4	76	0	1	51	1	3	0	1	0	14
13	1	0	8	0	18	0	0	1	0	1	2	1	0	9	58	1	0	10	0	0	5	0	1
15	0	0	1	0	3	0	9	0	1	0	7	2	1	0	1	72	1	2	2	2	1	0	0
12	0	8	1	3	0	1	1	1	0	0	1	0	0	0	15	1	0	84	0	0	0	1	0
9	1	0	0	1	1	0	1	5	1	0	0	0	1	0	0	0	4	2	0	94	1	0	1
20	0	0	1	3	21	3	3	5	0	1	1	20	1	0	1	3	0	2	0	0	84	2	0
19	2	5	1	0	3	1	0	0	0	0	19	1	2	2	1	0	0	1	3	0	0	78	0
24	0	1	0	1	0	0	2	20	0	0	0	0	1	0	0	0	0	7	0	0	1	0	29
4	4	1	0	16	0	0	2	0	13	2	0	1	5	0	0	1	0	0	24	0	6	0	0
5	4	1	2	23	4	54	45	3	1	5	5	2	3	3	2	2	0	1	10	0	1	2	0
6	21	0	8	3	1	1	4	4	0	0	1	35	44	1	1	1	0	1	0	10	0	0	0
18	2	0	0	0	0	0	6	6	47	1	1	1	0	32	0	19	0	1	2	0	0	0	0

Supplementary Table 9 – continued
[5]Human chromosomes

	Atlantic cod linkage groups																						
Chr[5]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	46	6	3	3	10	6	10	20	5	1	19	21	22	5	7	6	4	4	6	2	12	23	1
17	3	46	16	1	0	2	9	1	1	0	5	1	2	2	1	17	1	32	0	3	1	1	1
9	1	7	3	17	0	10	0	7	1	0	1	10	3	2	2	1	1	1	17	0	0	3	1
14	0	1	2	0	20	1	0	4	1	1	5	1	0	1	9	1	2	0	0	1	10	2	0
12	15	2	0	7	0	16	1	3	6	0	6	2	7	2	0	0	1	2	3	4	2	2	0
11	2	3	1	3	4	1	18	1	23	10	3	2	2	18	4	25	5	4	3	1	5	0	0
19	2	30	15	5	4	4	2	21	4	0	11	35	0	5	1	4	3	5	1	3	2	1	0
5	3	2	4	11	1	13	4	2	1	13	5	3	0	2	1	1	2	3	1	2	0	4	1
3	28	2	2	1	1	1	5	6	0	0	4	6	30	3	0	6	2	2	4	7	1	4	5
16	2	24	3	1	0	4	8	2	16	0	2	0	1	28	0	3	1	12	1	1	0	3	1
10	1	0	7	0	2	4	3	4	3	3	0	0	1	1	22	1	0	14	4	0	2	2	3
4	6	1	10	1	2	2	7	0	1	5	1	1	1	0	4	2	6	4	3	0	3	0	1
2	5	1	4	2	11	3	3	8	0	2	4	5	2	4	10	5	3	3	2	24	12	5	3
7	4	12	6	4	0	8	11	7	5	0	6	7	1	1	1	4	1	6	5	2	2	1	8
6	8	19	4	4	6	3	3	4	1	1	1	1	5	3	7	1	1	0	7	2	7	3	1
8	3	2	3	1	0	9	0	3	0	0	2	0	4	4	1	1	1	3	1	1	5	6	1
13	2	1	0	0	0	0	0	0	0	4	0	1	0	3	0	1	1	0	1	6	1	0	2
15	2	0	2	0	7	5	1	1	21	3	3	8	1	24	0	1	1	1	0	1	3	2	0
18	2	0	1	1	0	0	0	7	1	1	2	2	0	1	0	0	1	2	1	0	2	1	1
20	15	0	5	2	6	2	2	9	1	2	1	1	11	0	2	0	0	0	1	0	2	5	2
21	0	1	0	0	0	0	4	0	1	0	0	0	1	0	2	2	1	2	0	2	0	0	0
22	2	9	8	7	0	11	3	0	7	1	2	0	1	3	1	1	1	4	4	0	0	0	0
X	14	2	1	2	1	4	4	2	2	12	4	2	2	5	0	2	2	4	2	7	2	0	0
Y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0

Supplementary Table 10 Combined genotype frequencies of the *ß1* globin promoter and amino acid polymorphisms among eight Atlantic cod populations

Genotype[1]			Location[2]							
Pr	*ß1*-55	*ß1*-62	Bjørn-øya (40)	Båts-fjord (41)	Ma-langen (40)	Molde (40)	Helgo-land (36)	Katte-gat (75)	Born-holm (53)	Øland (36)
L	Val	Ala	0.700	0.805	0.600	0.150	0.139	0.107	0.925	0.944
*	*	*	0.100	0.122	0.175	0.225	0.250	0.387	0.055	0.028
S	Met	Lys			0.075	0.225	0.250	0.13		
L	*	*	0.050	0.024		0.100	0.195	0.094	0.020	
*	Met	Lys			0.025	0.125	0.139	0.120		
*	Val	*		0.049	0.075	0.100		0.040		0.028
*	Val	Ala	0.125			0.025				
S	*	*						0.040		
L	Val	*	0.025			0.025		0.013		
*	*	Lys						0.027		
S	Val	Lys			0.025			0.013		
S	*	Lys					0.027	0.013		
L	Met	Ala			0.025					
*	Met	*				0.025				
L	Met	Lys						0.013		

[1]Homozygote genotypes are scored as long (L) or short (S) for the promoter (Pr), Val or Met for the *ß1*-55, Ala or Lys for the *ß1*-62 polymorphisms and as heterozygote genotypes (*).
[2]Sample size is given in brackets. Bornholm and Øland are located in the Baltic Sea.

Supplementary Table 11 Presence or absence of selected immune-related genes based on the Atlantic cod annotation and additional manual curation

Gene name	Ensembl Identifier	Annotation evidence	Location[1]
	MHCI pathway		
MHCI U[2]	ENSGAUG00000001280	ENSGACP00000000184	GeneScaffold_20
MHCI U[2]	ENSGAUG00000002468	ENSGACP00000000197	GeneScaffold_22
MHCI Z	ENSGAUG00000000827	ENSGACP00000012843	GeneScaffold_1860
B_2-microglobulin	ENSGAUG00000002225	Q70XT2.1	GeneScaffold_2296
TAP1 (ABCB2)	ENSGAUG00000007212	ENSGACP00000017279	GeneScaffold_2525
TAP1 (ABCB2)	ENSGAUG00000015175	ENSGACP00000012286	GeneScaffold_1765
TAP2 (ABCB3)	ENSGAUG00000000616	ENSGACP00000005446	GeneScaffold_944
TAP2 (ABCB3)	ENSGAUG00000008942	ENSGACP00000005446	contig444738
TAP2 (ABCB3)	ENSGAUG00000001482	ENSGACP00000005446	GeneScaffold_4549
Tapasin (TAPBP)	ENSGAUG00000012873	ENSGACP00000000202	contig889828
PSME1 (PA28a)	ENSGAUG00000004020	ENSGACP00000002855	GeneScaffold_581
PSME2 (PA28b)	ENSGAUG00000018140	B5X6E1.1	GeneScaffold_1223
PSME3 (PA28g)	ENSGAUG00000002690	ENSGACP00000004854	GeneScaffold_879
PSME3 (PA28g)	ENSGAUG00000016730	ENSGACP00000007632	GeneScaffold_81
PSMB1	ENSGAUG00000012441	ENSGACP00000021315	GeneScaffold_299
PSMB2	ENSGAUG00000009518	ENSGACP00000003496	GeneScaffold_644
PSMB3	ENSGAUG00000012997	ENSGACP00000023984	GeneScaffold_3761
PSMB4	ENSGAUG00000008533	ENSGACP00000013097	GeneScaffold_1838
PSMB5	ENSGAUG00000002995	ENSGACP00000017491	GeneScaffold_2562
PSMB5[3]			scaffold06123
PSMB6	ENSGAUG00000013703	C1BKZ0.1	GeneScaffold_4633
PSMB6	ENSGAUG00000017774	ENSGACP00000001487	GeneScaffold_241
PSMB6	ENSGAUG00000005681	ENSGACP00000000189	GeneScaffold_21
PSMB7	ENSGAUG00000006798	ENSGACP00000024213	GeneScaffold_3735
PSMB8 (LMP7)	ENSGAUG00000004525	ENSGACP00000018806	GeneScaffold_405
PSMB9 (LMP2)	ENSGAUG00000005664	ENSGACP00000000188	GeneScaffold_21
PSMB10 (MECL)	ENSGAUG00000015564	ENSGACP00000024531	GeneScaffold_3567
PSMB10 (MECL)	ENSGAUG00000005692	A7KIL7.1	GeneScaffold_21
GranzymeB	ENSGAUG00000007191	B6DXC8.1	GeneScaffold_1983
GranzymeB	ENSGAUG00000008868	ENSGACP00000009787	GeneScaffold_2054
GranzymeB	ENSGAUG00000016171	ENSGACP00000009787	contig294500
Perforin	ENSGAUG00000013585	ENSGACP00000017940	GeneScaffold_1869
Perforin	ENSGAUG00000018988	ENSGACP00000009706	GeneScaffold_3733
FasL CD178	ENSGAUG00000019633	A4IH35.1	GeneScaffold_2997
Fas CD95[3]			scaffold02578
Erap1	ENSGAUG00000014414	ENSGACP00000018077	GeneScaffold_2079
Erap1	ENSGAUG00000002558	ENSGACP00000018077	GeneScaffold_2078
Erap2	ENSGAUG00000016556	ENSGACP00000012916	GeneScaffold_2659
Irap	ENSGAUG00000016581	ENSGACP00000012950	GeneScaffold_2659
UNC-93 B	ENSGAUG00000019535	Q9H1C4.1	GeneScaffold_3764
	MHCII pathway		
RFXANK	ENSGAUG00000000317	ENSGACP00000020897	GeneScaffold_1141
RFXANK	ENSGAUG00000007488	ENSGACP00000017336	GeneScaffold_1983
RFXAP	ENSGAUG00000020442	ENSGACP00000027184	GeneScaffold_1169
RFX5[b]			contig144014
RFX7	ENSGAUG00000004164	ENSGACP00000008064	GeneScaffold_1297
RFX7	ENSGAUG00000002477	ENSGACP00000020906	GeneScaffold_3322
CIITA	ENSGAUG00000015568	Q66X48.1	GeneScaffold_68
MHCII α[4]			
MHCII β[4]			
Invariant chain[4]			

Supplementary Table 11, continued

T-cell receptors (TcR)			
CD3e	ENSGAUG00000004354	A3RK72.1	GeneScaffold_1290
CD8a	ENSGAUG00000008940	ENSGACP00000011825	GeneScaffold_1681
CD8b	ENSGAUG00000008962	ENSGACP00000011833	GeneScaffold_1681
CD4[3]			contig102546
CD3 zeta[3]			scaffold05220
CD3 g/d[3]			scaffold01924
TCRβ	ENSGAUG00000000442	ENSGACP00000016425	GeneScaffold_3455
TCRβ	ENSGAUG00000000419	A6QQ27.1	GeneScaffold_3455
TCRα	ENSGAUG00000004898	A4JYR6.1	GeneScaffold_375
TCRδ	ENSGAUG00000009143	A4JYR6.1	GeneScaffold_374
TCRδ	ENSGAUG00000009139	A4JYQ7.1	GeneScaffold_374
TCRγ[3,5]			scaffold00324
AIRE[3]			contig122792
AICDA (AID)	ENSGAUG00000004114	ENSGACP00000013915	GeneScaffold_1960
RAG1	ENSGAUG00000003395	ENSGACP00000015155	GeneScaffold_2196
RAG1	ENSGAUG00000019135	ENSGACP00000015895	GeneScaffold_2324
RAG2	ENSGAUG00000003392	Q90XJ3.1	GeneScaffold_2196
Interleukins and interferons			
IL1B	ENSGAUG00000000345	ENSGACP00000019287	GeneScaffold_3003
IL6ST	ENSGAUG00000010501	ENSGACP00000024618	GeneScaffold_3744
IL8	ENSGAUG00000003767	ENSGACP00000002251	contig269450
IL8	ENSGAUG00000003939	ENSGACP00000002251	scaffold00254
IL8	ENSGAUG00000007034	ENSGACP00000002251	scaffold05973
IL8	ENSGAUG00000009294	ENSGACP00000002251	contig293714
IL8	ENSGAUG00000016625	ENSGACP00000002251	scaffold03038
IL8	ENSGAUG00000016716	ENSGACP00000002251	GeneScaffold_3084
IL8	ENSGAUG00000016719	B5U149.1	GeneScaffold_3084
IL8	ENSGAUG00000018176	ENSGACP00000002251	GeneScaffold_3076
IL10[3]			contig26533
IL12B	ENSGAUG00000015572	Q7SX69.1	GeneScaffold_3769
IL12B	ENSGAUG00000011082	ENSGACP00000027049	GeneScaffold_4322
IL12B	ENSGAUG00000007238	ENSGACP00000005652	GeneScaffold_3225
IL15[3,5]			contig730501/contig330188
IL17D	ENSGAUG00000015867	ENSGACP00000001806	GeneScaffold_2248
IL17A F1	ENSGAUG00000016229	Q5TKT0.1	GeneScaffold_542
IL22	ENSGAUG00000017088	C0MHM0.1	GeneScaffold_3987
IL22	ENSGAUG00000017108	C0MHM0.1	GeneScaffold_3987
IL2RG	ENSGAUG00000010773	ENSGACP00000027020	GeneScaffold_4322
IL2RG	ENSGAUG00000006115	ENSGACP00000027020	GeneScaffold_2546
IL2RB	ENSGAUG00000020552	ENSGACP00000010601	GeneScaffold_2887
IL4RA	ENSGAUG00000005712	B5X2F9.1	GeneScaffold_1668
IL8RB-Like	ENSGAUG00000019857	ENSGACP00000011585	GeneScaffold_2536
IL12RB2	ENSGAUG00000019306	Q6UAN1.1	GeneScaffold_1649
IL12RB2	ENSGAUG00000007599	C0H8Y1.1	GeneScaffold_2819
IL12RB2	ENSGAUG00000007606	Q6UAN1.1	GeneScaffold_2819
IL17RA	ENSGAUG00000001224	C0H963.1	GeneScaffold_3956
IL17RD	ENSGAUG00000011278	ENSGACP00000005075	GeneScaffold_904
FOXP3	ENSGAUG00000005614	ENSGACP00000016881	GeneScaffold_1197
TNFa	ENSGAUG00000013281	ENSGACP00000001821	GeneScaffold_351
TGFB	ENSGAUG00000018401	ENSGACP00000016928	GeneScaffold_816
IFNG[3]			contig127998
IPS1 (MAVS)	ENSGAUG00000015933	ENSGACP00000006456	GeneScaffold_2355
IKKG	ENSGAUG00000006193	ENSGACP00000017418	GeneScaffold_875
MYD88	ENSGAUG00000010938	ENSGACP00000004643	GeneScaffold_827

Supplementary Table 11, continued

Complement cascade			
C1qT4	ENSGAUG00000003385	ENSGACP00000022969	GeneScaffold_2870
C1qT4	ENSGAUG00000003764	ENSGACP00000015531	GeneScaffold_1479
C1qT5	ENSGAUG00000011784	ENSGACP00000011917	GeneScaffold_1739
C3	ENSGAUG00000010277	ENSGACP00000026212	GeneScaffold_4115
C3	ENSGAUG00000010740	ENSGACP00000024929	GeneScaffold_3850
C3	ENSGAUG00000012895	ENSGACP00000024919	GeneScaffold_3846
C3	ENSGAUG00000017964	ENSGACP00000024919	GeneScaffold_4051
C4	ENSGAUG00000009370	ENSGACP00000010542	GeneScaffold_752
C4	ENSGAUG00000009420	Q70TF5.1	GeneScaffold_752
C5[3]			contig84096
C6	ENSGAUG00000009978	ENSGACP00000023666	GeneScaffold_3719
C7	ENSGAUG00000007702	ENSGACP00000009161	GeneScaffold_1288
C7	ENSGAUG00000010005	ENSGACP00000023675	GeneScaffold_3719
C8	ENSGAUG00000017195	ENSGACP00000021571	GeneScaffold_4457
C8	ENSGAUG00000017166	ENSGACP00000021553	GeneScaffold_4457
C8	ENSGAUG00000008700	ENSGACP00000003228	GeneScaffold_648
C9	ENSGAUG00000008102	ENSGACP00000020928	GeneScaffold_1148
C9	ENSGAUG00000009172	ENSGACP00000020928	GeneScaffold_3215
B cells and APC's			
IgM	ENSGAUG00000014182	ENSGACP00000016893	GeneScaffold_2463
Igb	ENSGAUG00000013859	ENSGACP00000016918	scaffold07397
IgD[3]			contig124962
PTPRC (CD45)	ENSGAUG00000018523	ENSGACP00000010411	GeneScaffold_1180
CD79A	ENSGAUG00000018935	ENSGACP00000004224	GeneScaffold_4630
CD79B	ENSGAUG00000003007	ENSGACP00000004622	GeneScaffold_1533
CD79B	ENSGAUG00000020209	ENSGACP00000004622	GeneScaffold_1533
CD226	ENSGAUG00000008559	ENSGACP00000003543	GeneScaffold_661
CD40L	ENSGAUG00000004557	ENSGACP00000022786	GeneScaffold_3590
CD40	ENSGAUG00000004030	ENSGACP00000014752	GeneScaffold_4577
CD40	ENSGAUG00000003650	B6RCP8.1	GeneScaffold_1338
BLNK	ENSGAUG00000010236	C0H8Z7.1	GeneScaffold_709
IGBP1	ENSGAUG00000015071	Q6PI45.1	GeneScaffold_4043
Chemokines and receptors			
CXCR2	ENSGAUG00000019849	ENSGACP00000002998	GeneScaffold_1686
CXCR3	ENSGAUG00000016951	ENSGACP00000002339	GeneScaffold_4628
CXCR3	ENSGAUG00000016930	ENSGACP00000002323	GeneScaffold_4628
CXCR4	ENSGAUG00000009602	ENSGACP00000009641	GeneScaffold_2061
CXCR4	ENSGAUG00000012635	ENSGACP00000016367	GeneScaffold_2400
CCR5	ENSGAUG00000020143	ENSGACP00000019236	contig260864
CCR6	ENSGAUG00000003068	ENSGACP00000018230	GeneScaffold_2781
CCR7	ENSGAUG00000020472	ENSGACP00000023930	GeneScaffold_3761
CCR9	ENSGAUG00000020520	ENSGACP00000003192	GeneScaffold_383

Supplementary Table 11, continued

TLR-pathway			
TLR3	ENSGAUG00000000786	ENSGACP00000022288	GeneScaffold_4035
TLR7	ENSGAUG00000001675	ENSGACP00000005259	GeneScaffold_918
TLR8	ENSGAUG00000011526	B2GUH5.1	GeneScaffold_888
TLR8	ENSGAUG00000001639	ENSGACP00000005236	GeneScaffold_918
TLR8	ENSGAUG00000001689	Q3TM31.1	GeneScaffold_918
TLR9	ENSGAUG00000011256	B6D1N8.1	GeneScaffold_2185
TLR9	ENSGAUG00000003222	A8SZH4.1	GeneScaffold_4639
TLR9	ENSGAUG00000003161	Q3L274.1	GeneScaffold_4639
TLR9	ENSGAUG00000011244	ENSGACP00000013425	GeneScaffold_2185
TLR9	ENSGAUG00000003269	Q6Y1S0.1	GeneScaffold_4639
TLR14	ENSGAUG00000003793	ENSGACP00000002280	GeneScaffold_416
TLR16	ENSGAUG00000001699	Q4LDR7.1	GeneScaffold_3960
TLR21	ENSGAUG00000018200	B6EUP3.1	GeneScaffold_1988
TLR22	ENSGAUG00000000150	ENSGACP00000007198	GeneScaffold_1177
TLR22	ENSGAUG00000000143	ENSGACP00000007198	GeneScaffold_1177
TLR22	ENSGAUG00000000110	ENSGACP00000007198	scaffold03378
TLR22	ENSGAUG00000000152	ENSGACP00000007198	GeneScaffold_1177
TLR22	ENSGAUG00000010841	ENSGACP00000007198	GeneScaffold_1176

[1]Location follows the Ensembl annotation when a gene model is complete. If no gene model is present, BLAST searches provide a putative location.
[2]These are the only two U-lineage MHCI loci present in the assembly. The gene models contain one or more deletions, indicative of assembly errors.
[3]Reciprocal best BLAST hit using homologous protein sequences from rainbow trout (*Oncorhynchus mykiss*), brown trout (*Salmo trutta*), salmon (*Salmo salar*), zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*), tetraodon (*Tetraodon nigroviridis*), medaka (*Oryzias latipes*), catfish (*Ictalurus punctatus*), Atlantic cod (*Gadus morhua*), carp (*Cyprinus carpio*), halibut (*Paralichthys olivaceus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), frog (*Xenopus leavis*), black gibbon (*Hylobates concolor*) and human (*Homo sapiens*). If no full-length teleost sequence was available for a particular protein, homologs from the other vertebrate species were utilized (see also Supplementary Table 12). Query sequences were compared, in the order mentioned, to annotated genome transcripts, the genome assembly, cDNA assembly, and unassembled 454 sequencing reads and Illumina sequencing reads. Reciprocal BLAST results were obtained using the NCBI RefSeq database (Release 42).
[4]No reciprocal BLAST hit was obtained.
[5]Genomic location was found by aligning a cDNA contig (obtained by reciprocal BLAST hit) to the assembly.

Supplementary Table 12 Vertebrate homologs of MHCII, invariant chain and CD4

Name[1]	Genbank/Ensembl ID	Species	Common name
		MHCII alpha chain	
DAA	AF103003	*Ictalurid punctatus*	catfish
DBA	AF103005	*Ictalurid punctatus*	catfish
DMA	NM_001099353.1	*Gallus gallus*	chicken
DAA	L77086.1	*Salmo salar*	Atlantic salmon
DDA	DW557800	*Salmo salar*	Atlantic salmon
DCA	DW549478	*Salmo salar*	Atlantic salmon
DBA	EU008541	*Salmo salar*	Atlantic salmon
DAA	XM_001330976.1	*Danio rerio*	zebrafish
DAA	NM_001007205.1	*Danio rerio*	zebrafish
	ENSDART00000061132	*Danio rerio*	zebrafish
	AAH74082	*Danio rerio*	zebrafish
	ENSDART00000048448	*Danio rerio*	zebrafish
	ENSDART00000109439	*Danio rerio*	zebrafish
	ENSDART00000006898	*Danio rerio*	zebrafish
	ENSDART00000043525	*Danio rerio*	zebrafish
	ENSDART00000102847	*Danio rerio*	zebrafish
	ENSORLT00000000027	*Oryzias latipes*	medaka
	ENSORLT00000016021	*Oryzias latipes*	medaka
	ENSORLT00000024164	*Oryzias latipes*	medaka
	ENSORLT00000023575	*Oryzias latipes*	medaka
	AY997530.1	*Paralichthys olivaceus*	olive flounder
DAA	AY713945.1	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000025242	*Gasterosteus_aculeatus*	stickleback
DBA	ENSGACT00000004910	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000000425	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000000421	*Gasterosteus_aculeatus*	stickleback
DDA	ENSGACT00000000434	*Gasterosteus_aculeatus*	stickleback
	ENSTNIT00000008459	*Tetraodon nigroviridis*	Green spotted pufferfish
	ENSTNIT00000006494	*Tetraodon nigroviridis*	Green spotted pufferfish
	ENSTRUT00000002014	*Takifugu rubripes*	Japanese pufferfish
	ENSTRUT00000010379	*Takifugu rubripes*	Japanese pufferfish
	ENSTRUT00000044859	*Takifugu rubripes*	Japanese pufferfish
HLA-DMA	X76775	*Homo sapiens*	human
HLA-DOA	NP002119	*Homo sapiens*	human
HLA-DRA	NM_019111	*Homo sapiens*	human
HLA-DPA	NM_033554.2	*Homo sapiens*	human
HLA-DQA	AH002885.1	*Homo sapiens*	human

Supplementary Table 12, continued

		MHCII Beta chain	
DAB	IPU77598	*Ictalurid punctatus*	catfish
DMB	AB426143	*Gallus gallus*	chicken
DBB	EU008541	*Salmo salar*	Atlantic salmon
DAB	AJ438971.1	*Salmo salar*	Atlantic salmon
DCB	ENSDART00000098502	*Danio rerio*	zebrafish
	ENSDART00000098138	*Danio rerio*	zebrafish
	ENSDART00000075830	*Danio rerio*	zebrafish
	ENSDART00000077802	*Danio rerio*	zebrafish
DAB	ENSDARP00000070315	*Danio rerio*	zebrafish
	ENSDART00000040336	*Danio rerio*	zebrafish
	ENSDART00000000148	*Danio rerio*	zebrafish
	ENSDART00000097932	*Danio rerio*	zebrafish
	ENSDART00000108538	*Danio rerio*	zebrafish
	ENSDART00000000591	*Danio rerio*	zebrafish
	ENSDART00000099281	*Danio rerio*	zebrafish
	ENSORLT00000011498	*Oryzias latipes*	medaka
	ENSORLT00000000030	*Oryzias latipes*	medaka
	ENSORLT00000016052	*Oryzias latipes*	medaka
	ENSORLT00000024129	*Oryzias latipes*	medaka
	ENSORLT00000023543	*Oryzias latipes*	medaka
	AY848955.1	*Paralichthys olivaceus*	olive flounder
DAB	AF040760	*Xiphophorus maculatus*	platyfish
	AF194146.1	*Ginglymostoma cirratum*	nurse shark
	EB173954	*Hippoglossus hippoglossus*	Atlantic halibut
DAB	EU399184.1	*Epinephelus akaara*	Hong Kong grouper
	AM113471.1	*Dicentrarchus labrax*	European seabass
DAB	L33963.1	*Morone saxatilis*	striped bass
DAB	FJ372722.1	*Cynoglossus semilaevis*	sole
DAB	X95434.1	*Cyprinus carpio*	carp
DAB	AY713945.1	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000023783	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000025238	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000004851	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000000425	*Gasterosteus_aculeatus*	stickleback
	ENSGACT00000000437	*Gasterosteus_aculeatus*	stickleback
	ENSTNIT00000008460	*Tetraodon nigroviridis*	Green spotted pufferfish
	ENSTNIT00000002403	*Tetraodon nigroviridis*	Green spotted pufferfish
	ENSTRUT00000004765	*Takifugu rubripes*	Japanese pufferfish
	ENSTRUT00000044878	*Takifugu rubripes*	Japanese pufferfish
HLA-DMB	X76776	*Homo sapiens*	human
HLA-DOB	NP002120	*Homo sapiens*	human
HLA-DRB	NM_002124.1	*Homo sapiens*	human
HLA-DQB	M60028.1	*Homo sapiens*	human
HLA-DPB	NM_002121.4	*Homo sapiens*	human

Supplementary Table 12, continued

	Invariant chain (Ii, CD74)		
Ii-1	ENSDART00000026021	*Danio rerio*	zebrafish
Ii-2	NM_131372	*Danio rerio*	zebrafish
Ii-1	ENSORLT00000005537	*Oryzias latipes*	medaka
Ii-2	ENSORLP00000006180	*Oryzias latipes*	medaka
Ii-1	ENSTRUT00000013994	*Takifugu rubripes*	Japanese pufferfish
Ii-2	ENSTRUP0000001720	*Takifugu rubripes*	Japanese pufferfish
Ii-1	ENSGACT00000014324	*Gasterosteus_aculeatus*	stickleback
Ii-2	ENSGACT00000023857	*Gasterosteus_aculeatus*	stickleback
Ii-1	BT057821.1	*Salmo salar*	Atlantic salmon
Ii-2	BT049560.1	*Salmo salar*	Atlantic salmon
	CK424214.1	*Ictalurid punctatus*	catfish
	CK425769.1	*Ictalurid punctatus*	catfish
Ii-1	CV826589.1	*Paralichthys olivaceus*	olive flounder
Ii-2	CX725585.1	*Paralichthys olivaceus*	olive flounder
Ii-1	ENSTNIT00000010452	*Tetraodon nigroviridis*	Green spotted pufferfish
Ii-2	ENSTNIT00000017460	*Tetraodon nigroviridis*	Green spotted pufferfish
	AY597054.1	*Gallus gallus*	chicken
Ii-a	NM_001025159.1	*Homo sapiens*	human
Ii-b	NM_004355.2	*Homo sapiens*	human
Ii-c	NM_001025158.1	*Homo sapiens*	human
		CD4	
CD4-1	ENSDART00000103975	*Danio rerio*	zebrafish
CD4-2	NM_001135137	*Danio rerio*	zebrafish
CD4-1	AB274725.1 (fragment)	*Oryzias latipes*	medaka
CD4-2	ENSORLT00000015991	*Oryzias latipes*	medaka
CD4-1	NP_001123611	*Salmo salar*	Atlantic salmon
CD4-2	NP_001128603	*Salmo salar*	Atlantic salmon
CD4-1	NP_001072091.1	*Takifugu rubripes*	Japanese pufferfish
CD4-2	ENSTRUT00000027449	*Takifugu rubripes*	Japanese pufferfish
CD4	ENSGACP00000012984	*Gasterosteus_aculeatus*	stickleback
CD4-1	ABD93355.1	*Ictalurus punctatus*	catfish
CD4-2	DQ435304	*Ictalurus punctatus*	catfish
CD4	FJ185043	*Hippoglossus hippoglossus*	Atlantic halibut
CD4-1	ENSTNIP00000002054	*Tetraodon nigroviridis*	Green spotted pufferfish
CD4-2	ENSTNIP00000001516	*Tetraodon nigroviridis*	Green spotted pufferfish
CD4-3	ENSTNIP00000010843	*Tetraodon nigroviridis*	Green spotted pufferfish
CD4	DQ400124.1	*Cyprinus carpio*	common carp
CD4	AM849811.1	*Dicentrarchus labrax*	European seabass
CD4	NM_204649.1	*Gallus gallus*	chicken
CD4	NM_000616	*Homo sapiens*	human

[1]Not all genes have been assigned a name.

Supplementary Table 13 Teleost sequences used in MHCII alignment

Species	GenBank Id
Miichthys miiuy	HM236158.1
Larimichthys crocea	EF681865.1
Dicentrarchus labrax	AM113469.1
Morone saxatilis	L33967.1
Epinephelus akaara	EU399187.1
Epinephelus coioides	GU988723.1
Stizostedion vitreum	AY158837.1
Gasterosteus aculeatus	BT028490.1
Gasterosteus aculeatus	BT027207.1
Tetraodon nigroviridis	CR666800.3
Tetraodon nigroviridis	CR652355.3
Tetraodon nigroviridis	CR729735.3

Supplementary Table 14 Primers used in MHC qPCR

Assay	Primer name	5'- Primer sequence - 3'
MHC Class I[1]	Gaac_qPCR_A3_2F	ACCTCGGAGAGATCCTCCCCA
	Gaac_qPCR_A3_2R	CCTCGTCCACACCAGACAGC
	Gaac_qPCR_B2m_2F[3]	CCAAGAAAACACCCTCATCTGCCA
	Gaac_qPCR_B2m_2R[3]	CGCCAGCCCTGTTTGAAGGC
	Gamo_qPCR_A3_3F	CCCCAGTGGTGTGCCATGCT
	Gamo_qPCR_A3_3R	GGTCCCGTCGTGGTTGGGGA
	Gamo_qPCR_B2m_3F[3]	TCTGCCTGGTGCAAGCCTTCC
	Gamo_qPCR_B2m_3R[3]	TGGACAGGTGGAAGTGCCAGG
	Gamo_qPCR_topo_2F	ACGGCCCCAAACCACGTCAT
	Gamo_qPCR_topo_2R	AAGGTCAACCGGATGGGGCAC
	Hosa_qPCR_A3_3F	CCCTGGGCTTCTACCCTGCG
	Hosa_qPCR_A3_3R	CACAGCCGCCCACTTCTGGA
	Hosa_qPCR_B2m_1F	TGTCTGGGTTTCATCCATCCGACA
	Hosa_qPCR_B2m_1R	TCAGTGGGGGTGAATTCAGTGTAGT
MHC Class II[2]	Class_IIB_ex3_Forw1	CCATGYTGGTCTGCAGCGTST
	Class_IIB_ex3_Rev1.1	SCGTCTGCCAGCTCRTCAGT
	Class_IIB_ex3_Rev1.2	CGTCTGCCAGCTCRTCAGTC
	Class_IIB_ex3_Rev1.3	CTGGGCGTGTACTCCAGGTG
	Class_IIB_ex3_Rev1.4	GACCTGGGCGTGTACTCCAG
	Class_IIB_ex3_Forw2	CACCTGGAGTACACGCCCAG
	Class_IIB_ex3_Rev2	CCCAGGATCAGYCCBGASGCT
	Class_IIB_ex3_Forw3	ACACGCCCAGGTCYGGAGA
	Class_IIB_ex3_Rev3.1	GRCCCAGGATCAGYCCBGAGGCT
	Class_IIB_ex3_Rev3.2	CAGRCCCAGGATCAGYCCBGA

[1]Primers used for quantification of the MHCI α3 domain (A3), β-2-microglobulin (β2m) and topoisomerase (Topo) for stickleback (Gaac), Atlantic cod (Gamo) and human (Hosa) in qPCR experiments. Primers for topoisomerase were designed for Atlantic cod only.
[2]Primers tested for amplification of MHCII Beta exon 3, degenerate sites are underscored.
[3]Primers used in MHCII assay as positive controls.

Supplementary Table 15 Predicted locations of MHCI qPCR primers

Species	Primer target[1]	Hit ID	Hit location
Human	β2m	GI:224514848	15798250-15798391
	Class I like	GI:239740819	680-807
	Class I like	GI:239740716	677-804
	Class I like	GI:239740639	1582-1709
	Class I like	GI:89063264	725-852
	HLA-A	GI:269914190	708-835
	HLA-B	GI:170650640	740-867
	HLA-C	GI:52630341	693-820
	HLA-E	GI:301171456	803-930
	HLA-F	GI:149158701	801-928
	HLA-G	GI:269914083	864-991
	HLA-H	GI:269914105	839-966
	HLA-J	GI:209870098	734-861
	HLA-L	GI:240120085	572-699
Stickleback	β2m (Type 1)	ENSGACT00000025537	238-366
	Class I	ENSGACT00000000148	746-874
	Class I	ENSGACT00000000156	787-915
	Class I	ENSGACT00000000165	734-862
	Class I	ENSGACT00000000184	819-947
	Class I	ENSGACT00000000197	746-874
	Class I	ENSGACT00000002390	790-918
	Class I	ENSGACT00000002485	737-865
	Class I	ENSGACT00000002491	771-899
	Class I	ENSGACT00000002499	734-862
	Class I	ENSGACT00000002523	701-829
	Class I	ENSGACT00000002527	680-808
	Class I	ENSGACT00000002530	737-865
	Class I	ENSGACT00000002575	731-859

[1]Predicted locations of selected primer pairs for human and stickleback, and the location of the qPCR amplicon. Predictions based on Primer3 software output.

Supplementary Table 16 Teleost MHC Class I sequences

Transcript ID	Protein ID[1]	Amino acid length
ENSDART00000009689	ENSDARP00000020667	348
ENSDART00000016845	ENSDARP00000022240	333
ENSDART00000073378	ENSDARP00000067869	326
ENSDART00000073381	ENSDARP00000067872	332
ENSDART00000073382	ENSDARP00000067873	331
ENSDART00000073383	ENSDARP00000067874	328
ENSDART00000090386	ENSDARP00000084819	328
ENSDART00000104863	ENSDARP00000095633	417
ENSGACT00000000148	ENSGACP00000000148	340
ENSGACT00000000156	ENSGACP00000000156	347
ENSGACT00000000165	ENSGACP00000000165	336
ENSGACT00000000184	ENSGACP00000000184	363
ENSGACT00000000197	ENSGACP00000000198	363
ENSGACT00000002390	ENSGACP00000002383	344
ENSGACT00000002485	ENSGACP00000002477	359
ENSGACT00000002491	ENSGACP00000002483	364
ENSGACT00000002499	ENSGACP00000002491	341
ENSGACT00000002523	ENSGACP00000002515	324
ENSGACT00000002527	ENSGACP00000002519	340
ENSGACT00000002530	ENSGACP00000002522	332
ENSGACT00000002570	ENSGACP00000002561	416
ENSGACT00000002575	ENSGACP00000002566	337
GI|18128141	AAL59855	341
GI|18124183	AAL59856	320
GI|18124186	AAL59857	320
ENSORLT00000001202	ENSORLP00000001201	332
ENSORLT00000008080	ENSORLP00000008079	387
ENSORLT00000008255	ENSORLP00000008254	341
ENSORLT00000008514	ENSORLP00000008513	353
ENSORLT00000008540	ENSORLP00000008539	360
ENSORLT00000015541	ENSORLP00000015540	355
ENSORLT00000021463	ENSORLP00000021462	368
ENSORLT00000024360	ENSORLP00000024359	350
ENSORLT00000025228	ENSORLP00000025227	316
ENSTNIT00000000248	ENSTNIP00000002995	357
ENSTNIT00000003219	ENSTNIP00000002167	307
ENSTNIT00000003613	ENSTNIP00000000273	370
ENSTNIT00000013079	ENSTNIP00000012887	323

[1]Teleost MHCI sequences used for phylogenetic analysis. All annotated loci from Ensembl for zebrafish, stickleback, medaka and tetraodon were included. Three MHCI loci (NCBI) from nurse shark were included as outgroup. The MHCI sequences for these teleosts contain classical and non-classical MHCI loci.

Supplementary Table 17 Teleost TLR-sequences

TLR name[1]	Ensembl protein ID or contig location[2]
GAMO_3	ctg7180001512282:11168-7953
GAMO_7a	ctg7180001501447:14244-17064
GAMO_7b	ctg7180001501447:311-3140
GAMO_7c	ctg7180001501450:4209-6876
GAMO_7d	ctg7180001501448:3506-6442
GAMO_7e	ctg7180001501448:11577-14118
GAMO_8a	ctg7180001516087:10972-8015
GAMO_8b	ctg7180001501446:1-2117
GAMO_8c	ctg7180001516091:4501-1978
GAMO_8d	ctg7180001516095:3576-565
GAMO_8e	ctg7180001516095:9023-6101
GAMO_9a	ctg7180001462277:845-4057
GAMO_9b	ctg7180001531645:6095-3165
GAMO_9c	ctg7180001023750:4253-1044
GAMO_9d	ctg7180001531641:20261-17042
GAMO_9e	ctg7180001531641:11313-8014
GAMO_14	ctg7180001524304:3811-970
GAMO_18	ctg7180001491260:2152-950
GAMO_21	ctg7180001457326:6847-4103
GAMO_22a	ctg7180001524105:13891-20331
GAMO_22b	ctg7180001489875:9594-4251
GAMO_22c	ctg7180001524119:1877-6260
GAMO_22d	ctg7180001489879:9330-2874
GAMO_22e	ctg7180001495673:7823-1178
GAMO_22f	ctg7180001504607:18332-20740
GAMO_22g	ctg7180001525098:26335-23927
GAMO_22h	ctg7180001495677:2352-4759
DARE_1	ENSDARP00000063175
DARE_2	ENSDARP00000110559
DARE_3	ENSDARP00000014779
DARE_4ba	ENSDARP00000104680
DARE_4bb	ENSDARP00000028819
DARE_5b	ENSDARP00000068642
DARE_7	ENSDARP00000105671
DARE_8a	ENSDARP00000103011
DARE_8c	ENSDARP00000106955
DARE_8d	ENSDARP00000107821
DARE_9	ENSDARP00000105677
DARE_18	ENSDARP00000106955
DARE_19	ENSDARP00000094019
DARE_19b	ENSDARP00000115820
DARE_20	ENSDARP00000092183

DARE_20a	ENSDARP00000108970
DARE_21	ENSDARP00000098992
GAAC_2	ENSGACP00000024681
GAAC_3	ENSGACP00000022288
GAAC_5b	ENSGACP00000005783
GAAC_6	ENSGACP00000023718
GAAC_7	ENSGACP00000005259
GAAC_8	ENSGACP00000005236
GAAC_9	ENSGACP00000013425
GAAC_14	ENSGACP00000002280
GAAC_21a	ENSGACP00000012342
GAAC_21b	ENSGACP00000011097
GAAC_22	ENSGACP00000007198
HOSA_1	ENSP00000421259
HOSA_2	ENSP00000260010
HOSA_3	ENSP00000296795
HOSA_4	ENSP00000363089
HOSA_5	ENSP00000355846
HOSA_6	ENSP00000371376
HOSA_7	ENSP00000370034
HOSA_8	ENSP00000312082
HOSA_9	ENSP00000353874
HOSA_10	ENSP00000308925
ORLA_2	ENSORLP00000003159
ORLA_3	ENSORLP00000010277
ORLA_5a	ENSORLP00000020303
ORLA_6	ENSORLP00000005545
ORLA_8	ENSORLP00000022276
ORLA_9	ENSORLP00000011108
ORLA_14	ENSORLP00000019663
ORLA_18	ENSORLP00000015994
ORLA_21	ENSORLP00000016852
ORLA_22	ENSORLP00000025294
TARU_1	ENSTRUP00000025134
TARU_2	ENSTRUP00000007454
TARU_3	ENSTRUP00000012195
TARU_5	ENSTRUP00000011331
TARU_7	ENSTRUP00000020082
TARU_9	ENSTRUP00000019174
TARU_14	ENSTRUP00000001044
TARU_22	ENSTRUP00000015072
TARU_23	ENSTRUP00000010624
TENI_3	ENSTNIP00000009925
TENI_5b	ENSTNIP00000008738

TENI_6	ENSTNIP00000017225
TENI_7	ENSTNIP00000016967
TENI_9	ENSTNIP00000018058
TENI_22	ENSTNIP00000016626
TENI_23	ENSTNIP00000008068

[1]TLR names are identical to those used in Supplementary Figure 22. If the sequence has been given a TLR specific number as part of its gene name by Ensembl, the number in the name corresponds to this number. Otherwise, the name is a compound of the species name abbreviation and the clade specific TLR specification in Supplementary Figure 22. [2]Teleost TLR protein sequences used for phylogenetic analysis. All loci specifically annotated as TLR by Ensembl for zebrafish, stickleback, medaka, tetraodon, fugu and human were included. Additional sequences were also included that did not contain a TLR specification in their gene name (Supplementary Note 30). For Atlantic cod, we indicate sequence location through unique contig number and the start/stop of the entire TLR protein.

Supplementary Table 18 Number of sequenced 454 reads from cDNA of gadoids and salmon

Species	Tissue type[1]	
	Headkidney & spleen	Liver
Atlantic cod[2] (*G. morhua*)	423,470	
Haddock (*M. aeglefinus*)	486,068	
Whiting (*M. merlangus*)	420,702	
Burbot (*L. lota*)	555,534	
Atlantic salmon (*S. salar*)	132,151	60,665

[1]Samples were isolated and prepared from wild caught specimens following a similar methodology as described in Supplementary Note 12.
[2]Represents a specimen originating from a coastal Atlantic cod population from Lofoten.

Supplementary Table 19 Presence or absence of selected immune-related sequences in gadoids and salmon

Species[1]	Atlantic cod[2] (*G. morhua*)	Atlantic cod[3] (*G. morhua*)	Haddock (*M. aeglefinus*)	Whiting (*M. merlangus*)	Burbot (*L. lota*)	Atlantic salmon[4] (*S. salar*)
CD4	–[5]	–	–	–	–	+*
Ii	–	–	–	–	–	+*
MHCIIα	–	–	–	–	–	+*
MHCIIβ	–	–	–	–	–	+*
MHCI	+	+	+	+	+	+
β2m	+	+	+	+	+	+
CD8α	+	+	+	+	+	+**
CD8β	+	+	–*	+	+	+**
TCRα	+	+	+	+	+	+**
TCRβ	+	+	+	+	+	+**
ABCB2	+	+	+	+	+	+**
ABCB3	+	+	+	+	+	+**

[1]Presence was investigated in all species using assembled and unassembled cDNA sequence traces.
[2]Represents the sequenced specimen.
[3]Represents a specimen originating from a coastal Atlantic cod population from Lofoten.
[4]Represents a teleost with an empirically demonstrated functionality of the MHCII pathway
[5]Explanation of scores:
– no reciprocal BLAST hits using nine teleost, human and chicken homologs as query
–* no reciprocal BLAST hits using gadoid homologs as query
+ BLAST hit (maximum e-value10^{-10}) using Atlantic cod homologs as query
+* BLAST hit (maximum e-value10^{-50}) using nine teleost species, human and chicken homologs as query
+** BLAST hit (maximum e-value10^{-3}) using Atlantic salmon specific sequences (NCBI) as query.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21: 2104-2105.

Abrusan G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25: 1329-1330.

Aljanabi SM, Martinez I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res* 25: 4692-4693.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301-1310.

Bao Z, Eddy SR. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res* 12: 1269-1276.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27: 573-580.

Bergman CM, Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics* 8: 382-392.

Birney E, Clamp M, Durbin R. 2004. GeneWise and genomewise. *Genome Research* 14: 988-995.

Grosvik BE, Raae AJ. 1992. The genome size and the structure and content of ribosomal RNA genes in Atlantic cod (*Gadus morhua* L.). *Comp Biochem Physiol B* 101: 407-411.

Hardie DC, Hebert PDN. 2003. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome* 46: 683-706.

Hardie DC, Hebert PDN. 2004. Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1636-1646.

Hu Z, Bao J, Reecy J. 2008. CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online J Bioinformatics* 9: 108-112.

Hubert S, Higgins B, Borza T, Bowman S. 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics* 11: 191.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310-2314.

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR, Flanigan MJ, Edwards NJ, Bolanos R, Fasulo D, Halldorsson BV, Hannenhalli S, Turner R, Yooseph S, Lu F, Nusskern DR, Shue BC, Zheng XH, Zhong F, Delcher AL, Huson DH, Kravitz SA, Mouchard L, Reinert K, Remington KA, Clark AG, Waterman MS, Eichler EE, Adams MD, Hunkapiller MW, Myers EW, Venter JC. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci U S A* 101: 1916-1921.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.

Jakobsdóttir KB, Pardoe H, Magnússon Á, Björnsson H, Pampoulie C, Ruzzante DE, Marteinsdóttir G. 2011. Historical changes in genotypic frequencies at the Pantophysin locus in Atlantic cod (*Gadus morhua*) in Icelandic waters: evidence of fisheries-induced selection? *Evolutionary Applications*.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110: 462-467.

Karlen Y, McNair A, Perseguers S, Mazza C, Mermod N. 2007. Statistical significance of quantitative PCR. *BMC Bioinformatics* 8: 131.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin IT, Takeda H, Morishita S, Kohara Y. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447: 714-719.

Knutsen H, Olsen EM, Ciannelli L, Espeland SH, Knutsen JA, Simonsen JH, Skreslet S, Stenseth NC. 2007. Egg distribution, bottom topography and small-scale cod population structure in a coastal marine system. *Marine Ecology-Progress Series* 333: 249-255.

Kuhl H, Sarropoulou E, Tine M, Kotoulas G, Magoulas A, Reinhardt R. 2011. A Comparative BAC map for the gilthead sea bream (*Sparus aurata* L.). *J Biomed Biotechnol* 2011: 329025.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* 5: R12.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Lenz TL, Becker S. 2008. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci--implications for evolutionary analysis. *Gene* 427: 117-123.

Lerat E. 2009. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Huang Y, Yang G, Jiang Z, Qin N, Li L, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H. 2010. The sequence and *de novo* assembly of the giant panda genome. *Nature* 463: 311-317.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

Miller KM, Kaukinen KH, Schulze AD. 2002. Expansion and contraction of major histocompatibility complex genes: a teleostean example. *Immunogenetics* 53: 941-963.

Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L. 2010. Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. *Mol Biol Evol* 27: 2757-2774.

Osoegawa K, de Jong PJ, Frengen E, Ioannou PA. 2001. Construction of bacterial artificial chromosome (BAC/PAC) libraries. *Curr Protoc Mol Biol* Chapter 5: Unit 5 9.

Persson AC, Stet RJ, Pilstrom L. 1999. Characterization of MHC class I and beta(2)-microglobulin sequences in Atlantic cod reveals an unusually high number of expressed class I genes. *Immunogenetics* 50: 49-59.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1: i351-358.

Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, Harkins TT, Davidson WS. 2008. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* 9: 404.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

Rousset F. 2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol Ecol Resources* 8: 103-106.

Rozen S. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

Smit AFA, Hubley R. RepeatModeler Open-1.0.3 2008.
Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0.2 1996-2004.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456-463.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24: 1596-1599.
Wergeland HI, Jakobsen RA. 2001. A salmonid cell line (TO) for production of infectious salmon anaemia virus (ISAV). *Dis Aquat Organ* 44: 183-190.
Wetten OF, Nederbragt AJ, Wilson RC, Jakobsen KS, Edvardsen RB, Andersen O. 2010. Genomic organization and gene expression of the multiple globins in Atlantic cod: conservation of globin-flanking genes in chordates infers the origin of the vertebrate globin clusters. *BMC Evol Biol* 10: 315.
Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21: 1859-1875.
Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 18: 821-829.