Tørresen *et al. BMC Genomics* (2018) 19:240
https://doi.org/10.1186/s12864-018-4616-y

BMC Genomics

RESEARCH ARTICLE Open Access

Genomic architecture of haddock (*Melanogrammus aeglefinus)* shows expansions of innate immune genes and short tandem repeats

Ole K. Tørresen[1*], Marine S. O. Brieuc[1], Monica H. Solbakken[1], Elin Sørhus[2], Alexander J. Nederbragt[3,1], Kjetill S. Jakobsen[1], Sonnich Meier[2], Rolf B. Edvardsen[2] and Sissel Jentoft[1*]

Abstract

Background: Increased availability of genome assemblies for non-model organisms has resulted in invaluable biological and genomic insight into numerous vertebrates, including teleosts. Sequencing of the Atlantic cod (*Gadus morhua*) genome and the genomes of many of its relatives (Gadiformes) demonstrated a shared loss of the major histocompatibility complex (*MHC*) II genes 100 million years ago. An improved version of the Atlantic cod genome assembly shows an extreme density of tandem repeats compared to other vertebrate genome assemblies. Highly contiguous assemblies are therefore needed to further investigate the unusual immune system of the Gadiformes, and whether the high density of tandem repeats found in Atlantic cod is a shared trait in this group.

Results: Here, we have sequenced and assembled the genome of haddock (*Melanogrammus aeglefinus)* – a relative of Atlantic cod – using a combination of PacBio and Illumina reads. Comparative analyses reveal that the haddock genome contains an even higher density of tandem repeats outside and within protein coding sequences than Atlantic cod. Further, both species show an elevated number of tandem repeats in genes mainly involved in signal transduction compared to other teleosts. A characterization of the immune gene repertoire demonstrates a substantial expansion of *MCHI* in Atlantic cod compared to haddock. In contrast, the Toll-like receptors show a similar pattern of gene losses and expansions. For the NOD-like receptors (*NLRs*), another gene family associated with the innate immune system, we find a large expansion common to all teleosts, with possible lineage-specific expansions in zebrafish, stickleback and the codfishes.

Conclusions: The generation of a highly contiguous genome assembly of haddock revealed that the high density of short tandem repeats as well as expanded immune gene families is not unique to Atlantic cod – but possibly a feature common to all, or most, codfishes. A shared expansion of *NLR* genes in teleosts suggests that the *NLRs* have a more substantial role in the innate immunity of teleosts than other vertebrates. Moreover, we find that high copy number genes combined with variable genome assembly qualities may impede complete characterization of these genes, i.e. the number of *NLRs* in different teleost species might be underestimates.

Keywords: Haddock, Atlantic cod, STRs, Microsatellites, Genome assembly, NOD-like receptors

* Correspondence: o.k.torresen@ibv.uio.no; sissel.jentoft@ibv.uio.no
[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway
Full list of author information is available at the end of the article

Background

Recent advances in state-of-the-art genomic tools have resulted in a multitude of whole genome sequencing projects targeting non-model organisms. This has created a new understanding of the genomic basis of the biology of these species and their adaptation to the environment [1]. Examples include the adaptive radiation of African cichlids [2], adaptation to salinity in European sea bass and Atlantic herring [3, 4] and drastic morphological changes in pipefish and seahorses [5, 6], in addition to non-teleosts such as spotted gar and coelacanth, which aids in our understanding of the evolution of teleost fish [7, 8].

The species-rich order Gadiformes, i.e. codfishes and related species, comprises some of the most commercially important harvested fish in the world such as Alaska pollock (*Gadus chalcogrammus*), Atlantic cod (*Gadus morhua*), saithe (*Pollachius virens*) and haddock (*Melanogrammus aeglefinus)* [9, 10]. Recent reports have shown that this lineage has undergone dramatic evolutionary changes within its immune system compared to other jawed vertebrates, with a loss of the major histocompatibility complex (*MHC*) II genes in the lineage leading to the Gadiformes 105–85 million years ago [11, 12]. Additionally, other immune related genes have likely been lost prior to this event, e.g. the Toll-like receptor (*TLR*) 5151–147 million years ago and the Myxovirus resistance gene (*Mx*) 126–104 million years ago [13]. A detailed characterization of the *TLR* gene repertoire – membrane-bound receptors belonging to the pattern recognition receptors (PRRs) family and an important component of the innate immunity [14] – within the Gadiformes lineage revealed specific losses and several expansions [12, 15]. Some of these lineage-specific expansions, i.e. *TLR8*, *TLR22*, *TLR25* and in particular *TLR9,* were further correlated to the loss of *MHCII* and species latitudinal distributions [16]. An extreme expansion of *MHCI* genes – with more than 100 copies in some species – is another peculiarity of the immune system that Atlantic cod shares with many of the other gadiform species [11]. It has been suggested that some of these *MHCI* genes have taken on a more *MHCII*-like function through cross-presentation; i.e. compensating for the loss of the *MHCII* genes [17]. Taken together, these discoveries suggest that the loss of *MHCII* has fostered immunological innovation – through the altered *TLR* and *MHCI* gene repertoire – within the Gadiformes order.

Another important PRR family is the NOD-like receptors (NLR) class of proteins (also called NACHT-domain- and leucine-rich-repeat-containing receptors or nucleotide-binding domain and leucine-rich-repeat-containing receptors). These cytosolic receptors recognise microbial products and danger-associated molecular patterns [18]. The NLRs are a large class of intracellular immune receptors in animals [19]. Many species with a classic adaptive immune system contain relatively few *NLR* genes (around 20–30), such as mammals [18, 20]. Species without an adaptive immune system, such as cnidarians [21] and the purple sea urchin [22], contain large numbers of *NLRs* (up to 300). Investigations into the *NLRs* repertoire of teleosts indicate different numbers of *NLRs* in different species, e.g. a possible lineage-specific expansion in zebrafish [20].

The major impediment for creating highly contiguous genome assemblies in eukaryotes is the presence of repeated sequences [23]. For assemblies created solely from short Illumina reads (100–250 bp compared to 800–900 bp for Sanger) these repeated sequences can lead to fragmented assemblies missing important information, such as particular exons or whole genes [24]. With long-read sequencing (10,000 bp and longer as provided by PacBio and Oxford Nanopore), most of the repeats are likely to be spanned, and highly contiguous assemblies surpassing the earlier Sanger based assemblies in quality are possible [25–27]. Highly contiguous assemblies are a prerequisite for in-depth characterization and comparative studies of complex and multi-copy immune gene families (see [15]). Recently, a new version of the Atlantic cod genome assembly was generated by a combination of long read and conventional short read technologies, with substantial contiguity improvements compared to the previous version [28]. The improved assembly revealed an unusually high density of short tandem repeats (STRs, DNA motifs of 1–10 bp repeated in tandem) compared to other vertebrates [28]. STRs mutate at high rates [29], in humans from 10^{-8} to 10^{-2} mutations per locus per generation [30], and are located in about 4500 human genes [31]. Expression of about 2000 human genes is significantly associated with STR length variation in regulatory regions [32]. The Atlantic cod has about three times the density and frequency of STRs compared to humans, both in coding and non-coding regions [28]. Notably, this suggests that a substantially higher fraction of genes is associated with STRs in Atlantic cod compared to the human genome. These STRs might facilitate evolvability and rapid adaptation [33]. In humans, functional groups of genes such as “Transcription Factor and/or Development” and “Receptor and/or Membrane” have been identified as enriched in STRs [34]. Similar enrichment in functional groups have been identified in yeast [35], fruit fly [36] and in transcription and translation in plants and algae [37]. However, the degree to which Atlantic cod and other species of the Gadiformes share the same genomic distribution of these STRs within functional groups as in human and other species, is currently unknown and will require high-quality genome assemblies of additional gadiform species.

In this study, we have generated a highly contiguous genome assembly for haddock (*Melanogrammus aeglefinus*) using a combination of PacBio and Illumina reads. Our aim was to perform a comparative genomic analysis with the only other currently available highly contiguous gadiform genome assembly – that of Atlantic cod. The haddock assembly is comparable to the Atlantic cod assembly with regards to contiguity and gene content. Using this new assembly, we have further investigated the immune gene repertoire and the impact of STRs in Gadiformes. We show that ray-finned fish – including cod and haddock – are enriched for genes with STRs in functional groups (based on Gene Ontology) such as transcription factors. In addition, the codfishes (Atlantic cod and haddock) are significantly enriched for STRs in functional groups associated with signal transduction. Comparative analyses indicate a general expansion of the *NLR* genes in all teleosts, with possible lineage-specific expansions in zebrafish, stickleback and the codfishes.

Results

Assembly of the haddock genome

First, the different Illumina sequencing libraries were used to generate a genome assembly using the ALLPATHS-LG assembler [38] (see Methods). However, to obtain better assembly statistics (Table 1), we additionally generated an assembly using approximately 160× coverage of Illumina paired end reads and 20× coverage of PacBio reads with the Celera Assembler [39], resulting in a contig assembly (see Methods). All Illumina reads were mapped to the contig assembly with the Burrows-Wheeler Aligner (BWA) [40], and the scaffold module from String Graph Assembler (SGA) [41] was used to scaffold the contigs. To reduce gaps and to improve the accuracy of the consensus sequence, all Illumina reads were mapped to the scaffold assembly, and Pilon [42] was run to improve the contigs using high-coverage short-read information. Table 1 lists the statistics of the final assembly (also referred to as melAeg) and that of two assemblies from Tørresen et al. [28] for comparison. The melAeg assembly has shorter contigs and scaffolds than gadMor2, but approximately the same numbers of genes are found with CEGMA [43, 44] and BUSCO [45]. The GM_CA454PB assembly was one of the four assemblies combined to make gadMor2 [28], and it was created in a similar way to melAeg. It has similar contig and scaffold lengths, but fewer conserved genes were found by CEGMA and BUSCO.

Annotation and identifying orthologous genes

An iterative automatic annotation with MAKER [46, 47] using an Illumina based transcriptome of haddock created from reads sequenced by Sørhus et al. [48], and proteins from UniProt/SwissProt [49], annotated 96,576 gene models. InterProScan [50] was run on the predicted proteins of these, and gene names were allocated based on match with proteins in UniProt/SwissProt. We created a filtered set where all genes had an Annotation Edit Distance (AED) [51] of less than 0.5 (where 0.0 indicates perfect concordance between the gene model and evidence (mRNA and/or protein alignments), and 1.0 no concordance). This resulted in 27,437 gene models.

We used OrthoFinder [52] to create a catalogue of orthologous genes, inferring them based on the predicted proteins of different species. We included the following species from Ensembl r81: Amazon molly (*Poecilia formosa*), cave fish (*Astyanax mexicanus*), Atlantic cod (*Gadus morhua*; gadMor1), fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), platyfish (*Xiphophorus maculatus*), spotted gar (*Lepisosteus oculatus*), stickleback (*Gasterosteus aculeatus*), tetraodon (*Tetraodon nigroviridis*), tilapia (*Oreochromis niloticus*) and zebrafish (*Danio rerio*), in addition to haddock and the most recent Atlantic cod genome assembly (gadMor2). For each gene, only the longest protein isoform was used. 281,838 proteins were placed into 17,519 orthogroups, with 20,661 proteins without a match. Cod and haddock have 11,500 groups in common (at least one protein from each species). See Additional file 1: Table S1 for the number of orthogroups shared between the other species-pairs.

Table 1 Genome assembly statistics for haddock (melAeg) compared with an ALLPATHS-LG assembly and two assemblies of Atlantic cod, one draft based on PacBio and 454 reads (GM_CA454PB) and the final gadMor2 assembly

	melAeg	ALLPATHS-LG	GM_CA454PB	gadMor2
Length assembly (Mbp)	653	592	681	644
N50 scaffold (kbp)	209	169	272	1150
N50 contig (kbp)	78	4.4	95	116
CEGMA complete (% of 458 genes)	439 (96%)	428 (93%)	431 (94%)	435 (95%)
BUSCO single	4041 (88%)[a]	3562 (78%)[a]	3819 (83%)[a]	4160 (91%)[a]
BUSCO duplicated	128 (2.8%)[a]	92 (2.0%)[a]	117 (2.6%)[a]	127 (2.8%)[a]
BUSCO fragmented	203 (4.4%)[a]	407 (8.8%)[a]	359 (7.8%)[a]	139 (3.0%)[a]
BUSCO missing	212 (4.6%)[a]	523 (11%)[a]	289 (6.3%)[a]	158 (3.4%)[a]

[a] % of 4584 genes

Genetic variation and historic effective population size

To be able to compare the heterozygosity rate between haddock and cod, we mapped the Illumina reads of the two species from Malmstrøm et al. [11] against the assemblies with BWA [40], and called SNPs (single nucleotide polymorphisms), MNPs (multi-nucleotide polymorphisms), indels (insertions and deletions) and complex regions (composite insertion and substitution events) with FreeBayes [53]. Haddock had 40% more SNPs than cod (gadMor 2), with even larger differences in MNPs, indels and complex variants (Table 2).

While we have investigated only one individual per species, in general there is a correlation between nucleotide diversity of one individual and effective population size [54]. We used Pairwise Sequentially Markovian Coalescent (PSMC) [55] to infer the historic effective population size for the two species (Fig. 1). We used a generation time of 10 years for cod and 6 years for haddock [56] with mutation rates derived from the phylogeny used in Malmstrøm et al. (2016) [11]. From this we found that haddock has an approximately 2.5 times larger historic effective population size than cod (Fig. 1).

The *TLR* repertoire

Cod and haddock in general display the same *TLR* repertoire (Table 3). There is a difference of one or two gene copies for the cod assembly compared to what has been reported previously [15]. Our search criteria were quite strict, and the underlying assemblies were different (GM_CA454PB in [15], gadMor2 here), so some discrepancy can be expected.

Thirty-six full-length *TLRs* were identified for cod, whereas 28 were identified for haddock (Table 3). For both species, *TLRs* 1/6, 2, 4, 5, 21beta and 26 were not present. The gene numbers for most of the *TLRs* (*TLR* 3, 7, 9, 14, 21, 22, 23 and 25) were similar between both species. In contrast, cod had a significantly higher number of *TLR22* (10) than haddock (5).

The *MHCI* repertoire

The number of *MHCI* loci has previously been characterized in cod, using both qPCR and read-depth comparisons, with 80–100 and ~ 70 copies were estimated, respectively [11, 12]. By using read-depth comparisons for haddock, ~ 30 copies were calculated for this species [11]. Only two copies of *MHCI* were found in the first version of the cod genome assembly (gadMor1) [12]. We used the new assemblies of cod and haddock to investigate the number of copies of *MHCI.*

Table 2 Number of variants called for the assemblies of haddock and cod. In parenthesis the number of variants are given per bp, i.e. as nucleotide diversity

	Haddock	Cod
SNPs	3,552,609 (5.4×10^{-3})	2,506,699 (3.9×10^{-3})
MNPs	127,929 (0.2×10^{-3})	88,869 (0.1×10^{-3})
indels	1,013,087 (1.6×10^{-3})	608,828 (0.9×10^{-3})
complex	300,678 (0.5×10^{-3})	173,128 (0.3×10^{-3})

We inferred the presence of *MHCI* based on the occurrence of the three alpha domains of MHCI, including the most conserved alpha-3 domain. We found 13 regions with all three exons in cod, and 10 such regions in haddock. One significant difference between the two species was the number of occurrences of isolated alpha domains, suggesting potentially more copies of *MHCI* in cod (Table 4). Because these genes occur in multiple copies within the genome, the genome assembler might consider them as repeats [23], potentially resulting in fragmented assembly of these genes. We found up to 20 copies of *MHCI* (sum of all hits) in haddock, and 53 in cod, i.e., 66% and 76% of the previous estimated number of *MHCI* copies in haddock and cod, respectively [11].

Celera Assembler, the assembler used for assembling melAeg and GM_CA454PB, outputs so-called unitigs in addition to outputting contigs and scaffolds. Unitigs are sequences that are either unique in the genome or are collapsed repeated sequence. These are incorporated into contigs based on different rules (e.g., likelihood of being a repeat). Often, the contigs only contain a subset of the unitigs, and therefore could contain fewer genes. We translated the unitigs assemblies of melAeg and GM_CA454PB into all six reading frames with transeq [57] and searched these with the MHCI PFAM [58] domain PF00129, consisting of alpha 1 and 2, using HMMER [59]. For cod and haddock, the domain spans two exons, thus we counted occurrences of the first and last part of the profile found in the assemblies (Table 4). We found 27 copies of the first part of the domain and 30 copies of the last part in haddock and 69 and 70, respectively, in cod, approximately the same as in Malmstrøm et al. (2016) [11]. It is likely that some of these are collapsed because of the repeated nature of *MHCI* genes.

Expansion of *NLRs* in teleosts

The zebrafish has a lineage-specific expansion of the *NLRs* [60], but it is unclear how many copies are found in other teleost genome assemblies. We investigated the *NLRs* with several approaches. First, we ran InterProScan [50] on the longest protein per gene to annotate protein domains. We parsed the output and counted occurrences of the PFAM [58] domains PF05729 (NACHT domain) and PF14484 (Fish-specific NACHT associated domain, FISNA) (Fig. 2). Second, we translated the assemblies into all six reading frames with transeq [57] and used these to search for the NACHT and FISNA domains using HMMER [59]. For all species, the number of domains identified was substantially elevated when scrutinizing the assemblies compared

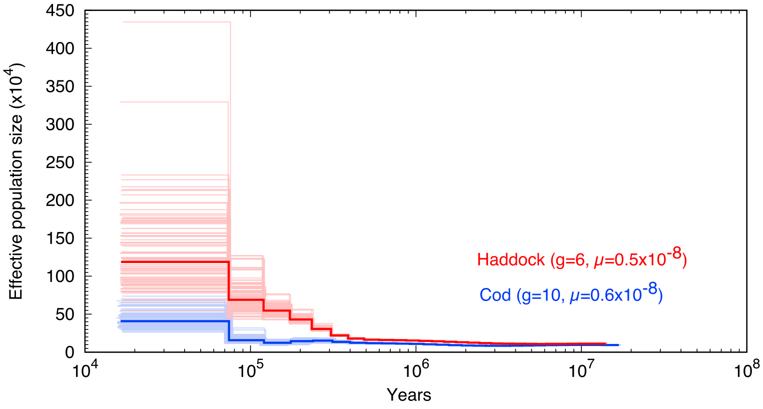


Fig. 1 The historic effective population sizes in cod and haddock. The analysis also includes the time before the two species split, as inferred by PSMC. Haddock is marked in red and cod in blue. Each analysis has been run with 100 bootstrap replicates, shown as pale versions of the main color. The time-span ranges from approximately 20 million to 20,000 years ago

to the predicted proteins (Fig. 2). For example, in platyfish the number of NACHT domains increased from 29 to 120. The reported numbers show a large variation in copy number between the different species (Fig. 2), with large difference between relatively closely related species, such as tetraodon and fugu, or cod and haddock, where there are three times as many copies in cod compared to haddock.

For the species with contigs/scaffolds placed into either linkage groups or chromosomes (cod, stickleback, zebrafish, spotted gar, medaka and tetraodon) we counted the number of genes where the relevant domains were found in either placed (i.e. in the linkage map) or unplaced sequences (Fig. 2, Additional file 1: Tables S5-S10). We found that many of the sequences with these kinds of domains are unplaced, as previously reported [20, 60]. While zebrafish has a majority of domains in placed sequences, most sequences in stickleback with FISNA and NACHT domains are not placed. About half the sequences are placed in cod, while most sequences are placed in the other species.

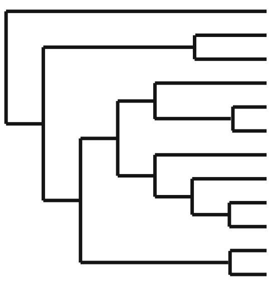
There are multiple reasons for a genome to not assemble properly, but repeated sequence is one of the most influential [23]. Genes occurring in multiple copies such as *NLRs* are indistinguishable from any other repeated sequence for the assembler. One consequence of this is that some of these unplaced contigs/scaffolds would have higher coverage in reads than average since they basically are collapsed repeats. For haddock and cod we have sequencing read data available, and we estimated and plotted the average coverage for all sequences with the FISNA domain (Fig. 3). Many of the sequences shorter than 100,000 bp show a higher than average

Table 3 Number of full-length *TLR* genes found in the haddock and cod assemblies. Additional incomplete copies (≥60% of the entire gene) are indicated in parenthesis

TLR gene	Haddock	Cod
TLR1/6	0	0
TLR2	0	0
TLR3	1	1
TLR4	0	0
TLR5	0	0
TLR7	1(2)	3
TLR8	8(1)	9
TLR9	5(1)	4(1)
TLR14	1	1
TLR21	2	1(1)
TLR21beta	0	0
TLR22	5(1)	10
TLR23	1	2
TLR25	4	5
TLR26	0	0

Table 4 The number of *MHCI* found in the haddock and cod assemblies based on different criteria. The BLAST-based reports open reading frames for the hits in the final assemblies, while the PFAM domain-based report the number of domains found in the unitig assemblies that underlie the final assemblies

	domain	Haddock	Cod
BLAST-based search (in melAeg and gadMor2)	alpha 1 + 2 + 3	10	13
	alpha 1	2	13
	alpha 2	0	7
	alpha 3	3	16
	alpha 1 + 2	2	0
	alpha 2 + 3	3	4
PFAM domain based search in unitig assemblies	first part (alpha 1)	30	69
	last part (alpha 2)	27	70



	Predicted proteins		Genome assembly	
	# NACHT	# FISNA	# NACHT (placed)	# FISNA (placed)
Spotted gar	34	15	32 (19)	16 (2)
Cavefish	97	90	107 (NA)	115 (NA)
Zebrafish	348	335	420 (380)	401 (361)
Stickleback	87	234	320 (52)	326 (46)
Fugu	42	38	76 (NA)	106 (NA)
Tetraodon	19	7	36 (10)	46 (10)
Tilapia	112	107	208 (NA)	266 (NA)
Medaka	26	18	58 (39)	93 (53)
Amazon molly	94	82	120 (NA)	111 (NA)
Platyfish	29	14	120 (NA)	111 (NA)
Atlantic cod	133	137	178 (93)	191 (90)
Haddock	36	41	59 (NA)	71 (NA)

Fig. 2 NACHT and FISNA domains content in predicted proteins and genome assemblies for the different species. HMMER hits had to be > 75% of the length of the domain to be reported here. Some species have scaffolds ordered and organized into chromosomes/linkage groups, i.e., placed. For these species the number of domains found in placed scaffolds are also reported. The phylogenetic relationship between the species is based on Malmstrøm et al. [9]. NA: Not applicable

coverage. This is especially the case for those sequences around 10,000 bp, and indicates that these contain multiple copies of the FISNA domain, i.e. these contain collapsed copies.

Due to differences in the assembly strategy, the haddock assembly contains fewer short contigs than the cod assembly (Additional file 1: Note S1). We investigated the unitig assemblies for cod and haddock with the NACHT and FISNA domains, with the same approach as used for *MHCI* for unitig assemblies (Table 5). This approach reports around 600 copies of each of the domains in both species. The NACHT domain is longer (166 aa) than the FISNA domain (72 aa), and while the total number of hits is similar between the two domains, there are significantly fewer NACHT domains found at > 75% of the domain length. The short hits for the NACHT domain are predominantly found on unitigs shorter than 500 bp, suggesting that these are collapsed.

Investigating the STR content of the haddock genome assembly

We investigated the amount of short tandem repeats (STRs) in the haddock genome assembly, compared to cod and other ray-finned fishes. We used Phobos [61] to annotate all STRs with an unit size of 1–10 bp. Haddock has an even higher density of STRs in its genome assembly compared to cod, 96,364 bp/Mbp in haddock and 80,706 bp/Mbp in cod (Fig. 4a). The amino acid coding parts of the genome also contain a high proportion of STRs, 25,639 bp/Mbp in haddock and 16,501 bp/Mbp in cod. This mostly consists of dinucleotide repeats, but both cod and haddock have approximately 6000 bp/Mbp of trinucleotide STRs in protein coding regions, compared to 530 bp/Mbp in medaka, and up to 934 bp/Mbp in zebrafish with the other fishes harboring intermediate amounts (Fig. 4b). Cod and haddock also have higher frequencies (loci/Mbp) of STRs in the assemblies (Fig. 4c and Additional file 1: Table S2), and in the protein coding regions (Fig. 4d). By using the overlap between annotated STRs and genes, we also report the number of genes with one or more STR for these species (Additional file 1: Table S3).

For haddock and cod, we were also able to find indels (called by FreeBayes) and STRs in protein-coding regions, and where these structural variants overlap. We found STRs of all unit sizes in the protein coding regions (Fig. 4d), but those STRs with unit sizes that do not create frame shifts, such as tri-, hexa- and enneanucleotides, are most interesting from a functional perspective. Of these, the vast majority are trinucleotides, and we restricted our analysis to these. We found 581 genes with an indel of size 3 in a trinucleotide repeat in haddock (2.1%) and 660 genes in cod (2.9%), i.e. these are heterozygous in these two individuals.

Between-species comparisons of STR enrichment in genes

Cod and haddock have a much larger proportion of their protein coding sequence in dinucleotide and trinucleotide STRs compared to other species (Fig. 4). In the process of annotating a genome, many genes are assigned a gene ontology term (GO term), describing the processes the protein encoded by that gene is involved in. We wanted to investigate if genes with STRs are randomly spread across different GO groups, or if some GO groups in some species are enriched for genes with STRs. Fisher's exact test was used to perform pairwise comparisons of the number of genes with STRs and the number of genes without STRs between each species (Fig. 5 for examples, Additional file 2: Figure S1 and Additional file 1: Table S4 for details). Of the 2748 GO terms in the dataset, there are significant differences between species in 74 GO groups after correcting for multiple testing (false discovery rate with Benjamini/Yekutieli). For many of these, haddock and cod differ significantly from all other species, but not from each other (Additional file 1: Table S4). These include

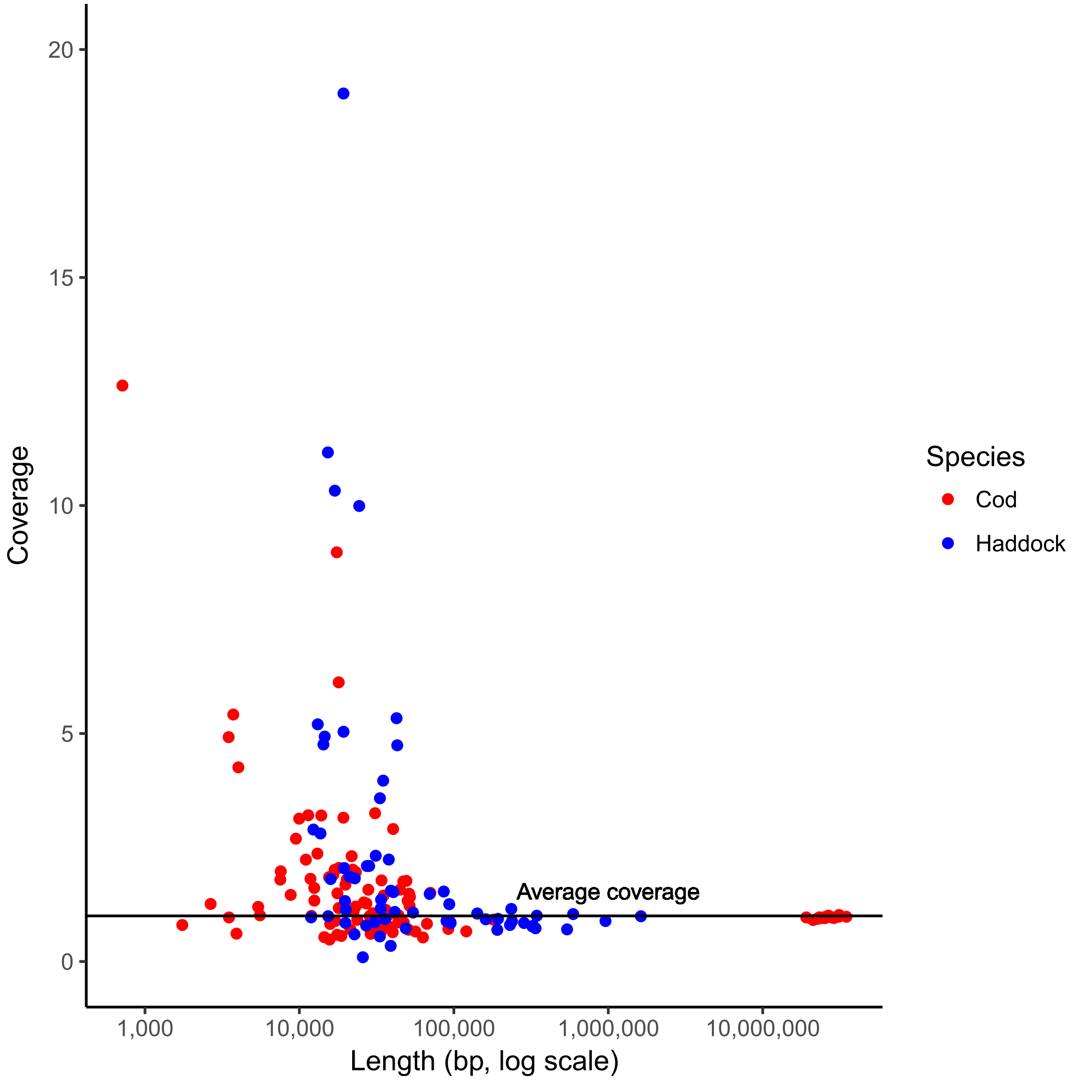


Fig. 3 Relationship between length and coverage of reads for sequences harboring the FISNA domain. Coverage has been normalized for each species by dividing the coverage for each sequence with the average for that species. The average lengths of genes with the FISNA domain is 17 kbp in cod and 14 kbp in haddock, and the increased coverage in sequences about this length might indicate that there are multiple, very similar regions with these genes in the two species. The cod sequences larger than 10 Mbp represent the linkage groups. Cod is plotted with red and haddock in blue. The x-axis is log(10)-transformed since the sequences span from 700 bp to more than 20 Mbp

protein kinase activity (GO:0004672), G-protein coupled receptor activity (GO:0004930), signal transduction (GO:0007165), metabolic process (GO:0008152) and transmembrane transport (GO:0055085).

Within-species comparisons of STR enrichment within genes

To investigate enrichment and purification (under-representation) of STRs in GO terms, we used goatools.

[62] (Fig. 6, Additional file 3: Figure S2). We corrected for multiple testing. For some terms, both cod and haddock are enriched, whereas this is not the case in the other species. These are cation channel activity (GO:0005261), regulation of signal transduction (GO:0009966), regulation of cell communication (GO:0010646), regulation of signaling (GO:0023051), regulation of Rho protein signal transduction (GO:0035023), regulation of Ras protein signal transduction (GO:0046578), regulation of response to stimulus (GO:0048583), regulation of small GTPase mediated signal transduction (GO:0051056), regulation of intracellular signal transduction (GO:1902531). These are mainly in the hierarchy above regulation of Rho protein signal transduction (GO:0035023), as well as cation channel activity (GO:0005261).

Discussion

A highly contiguous genome assembly for haddock

Here we have taken advantage of long and short read technologies to produce an annotated and highly

Table 5 The number of hits for NACHT and FISNA domains in the unitig assemblies for cod and haddock, as a proxy for number of *NLR* genes. Substantially more hits are found in the unitigs that in the contigs of the final assemblies, indicating that many of the unitigs are not included, possibly because they are categorized as repetitive sequence

Domain		Haddock	Cod
NACHT	all	613	656
	> 50% domain length	224	264
	> 75% domain length	121	140
	> 75% domain length, with stop codons	46	51
FISNA	all	611	552
	> 50% domain length	553	505
	> 75% domain length	384	359
	> 75% domain length, with stop codons	75	107

contiguous assembly of the haddock genome, with comparable gene content and assembly statistics to the recently released Atlantic cod genome assembly [28] (Table 1). The genic completeness of the assembly is high, as seen by the BUSCO score, where > 90% of the 4584 genes are found complete (Table 1). PacBio reads span more repeated regions than Illumina reads, and the contig N50 is therefore longer for the haddock assembly than other fishes sequenced with only Illumina reads, for instance the Asian arowana [63] and the seahorse [5]. With the increased affordability, availability and usage of such long-read sequencing technologies as PacBio [64] and Oxford Nanopore [65] reads, more complete assemblies for diverse species are likely to arrive in near future.

Increased number of tandem repeats in codfishes

Several studies have shown the Atlantic cod genome has a high STR content [66–68]. The first version of the cod genome assembly [12] was fragmented, and STRs have recently been identified as the main factor causing this fragmentation [28]. Since STRs have a high mutation rate, their presence in genes might disrupt normal gene product function, as seen for the multitude of human diseases due to large expansions in STRs [69]. Surprisingly, while both cod and haddock have a high density and frequency of STRs in the assembly overall, they also have a substantial amount of STRs in protein coding regions compared to other ray-finned fish (Fig. 4). STRs shrink and expand by DNA polymerase slippage or recombination [29], but a repeated motif has to be present for this to happen. A short tandem repeat might be created by a mutation (changing ATAG to ATAT), or as the result of transposable element activity [70]. Further work is needed to investigate the basis for the high STR

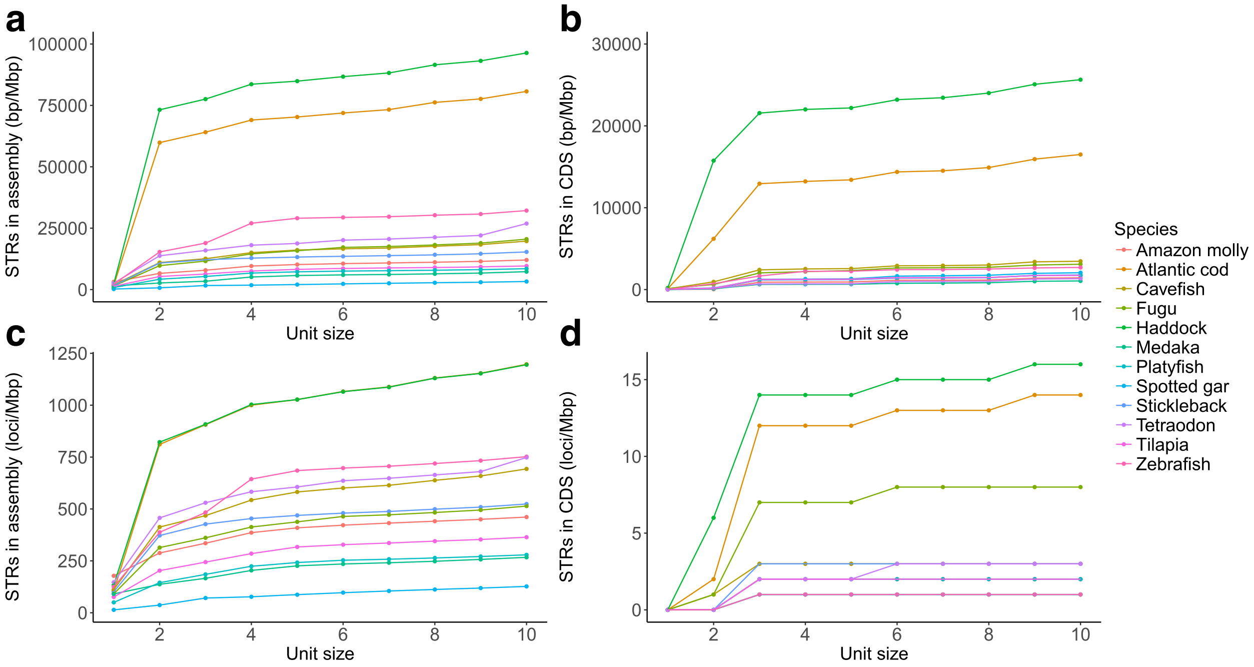


Fig. 4 Cumulative plot of the density (bp/Mbp) and frequency (loci/Mbp) of short tandem repeats (STRs). Shown is the STR content per unit size in the whole assembly and CDS for different teleosts. Most of the STR contents in the whole assembly in cod and haddock are dinucleotide repeats, but there are about equal amounts of dinucleotide and trinucleotide repeats in coding sequence. **a** Density of STRs in the genome assembly (bp/Mbp). **b** Density of STRs in protein-coding regions (bp/Mbp). **c** Frequency of STRs in the genome assembly (loci/Mbp). **d** Frequency of STRs in the protein-coding regions (loci/Mbp)

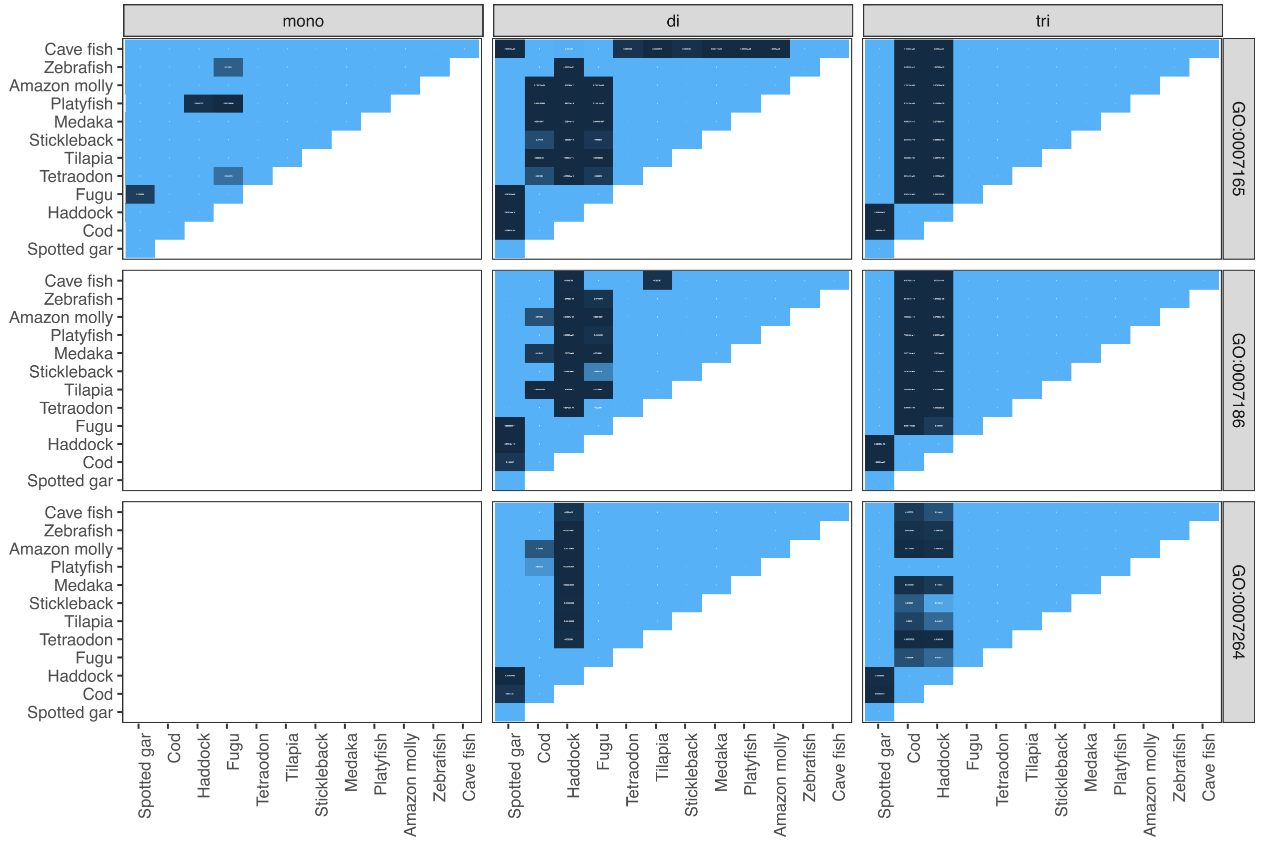


Fig. 5 Pairwise Fisher's exact test for some gene ontology groups and for some unit sizes. See Additional file 2: Figure S1 for the entire figure with 74 GO groups and unit sizes 1–10 bp, and Additional file 1: Table S4 for the GO groups where haddock and cod differ significantly from the other species. Shown here are GO:0007165 (signal transduction), GO:0007186 (G-protein coupled receptor signaling pathway) and GO:0007264 (small GTPase mediated signal transduction) for tandem repeats in 1–3 bp unit sizes. In the white panels and the white and light blue areas there are no significant differences, but in the dark blue areas there are significant differences between two species. For GO:0007165 and GO:0007186 there is a significant difference ($P < 0.05$) between cod and haddock and the other species, but not between cod and haddock, nor between cod and cave fish. For GO:0007264, this pattern is less apparent

content in Atlantic cod and haddock and in codfishes in general.

STRs are present in almost twice as many genes in cod and haddock compared to the other ray-finned fishes (Additional file 1: Table S3). Specifically, in around 8000 genes in codfishes compared to 1500–4000 in the other species. This is almost twice as many as in humans (4500) [31]. In humans, genes connected to processes such as transcriptional regulation, chromatin remodeling, morphogenesis, and neurogenesis have been found enriched for STRs [34, 71]. Similar enrichment has been found in other species, such as yeast [35], fruit fly [36] and plants and algae [37]. In the fish species investigated here, there is enrichment in genes with STRs in functional (Gene Ontology) groups primarily concerned with transcription, similar to previous studies [35–37] (Additional file 3: Figure S2). One example is the transcriptional regulator Ssn6 in yeast, where increased length of a polyglutamine tract (encoded by a STR), was positively correlated with increased expression of some target genes, and negatively correlated with others [72]. Haddock and cod have significantly larger proportions of genes with STRs in GO groups associated with genes encoding proteins involved in signal transduction compared to the other species. These GO groups contain a higher proportion of genes with STRs than expected with comparing GO groups per species. This is also true when comparing GO groups between species. Many of these functional groups are connected to small GTP-binding proteins such as regulation of Rho protein signal transduction (GO:0035023), regulation of Ras protein signal transduction (GO:0046578), and regulation of small GTPase mediated signal transduction (GO: 0051056). The small GTP-binding proteins are involved in regulation of processes such as gene expression, cytoskeletal reorganization, intracellular vesicle trafficking and cytokinesis [73, 74]. The regulation of the activity of small GTPases are mainly performed by GTPase-

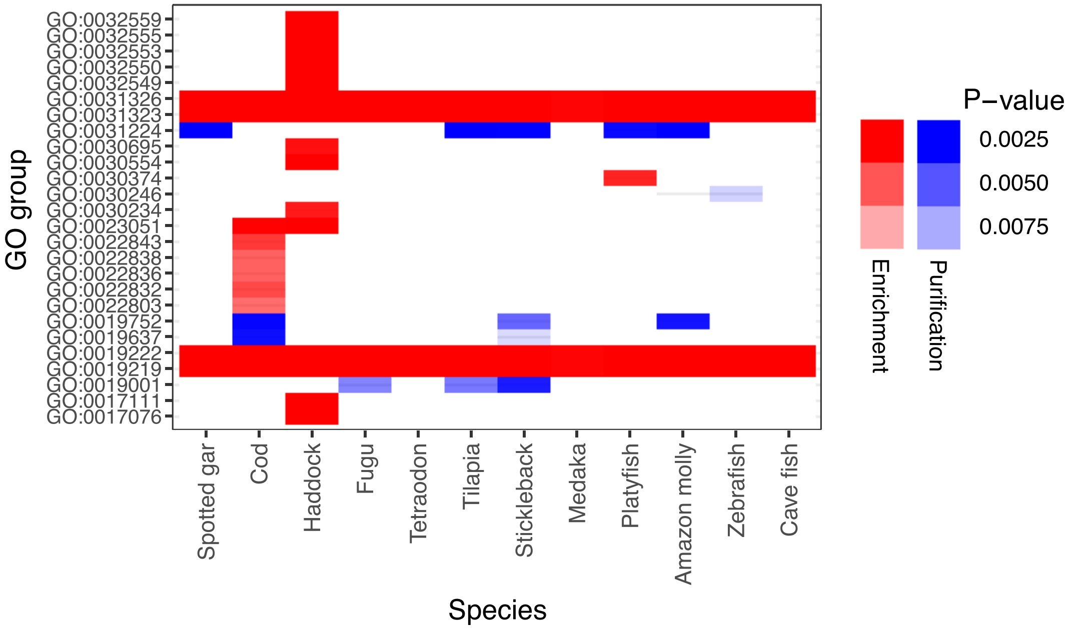


Fig. 6 An example of gene ontology terms significantly enriched for genes with trinucleotide tandem repeats in different species. Trinucleotide tandem repeats are repeats that can vary in number of repeat units without causing frameshifts in the protein. Only tests with $P < 0.01$ are colored. Red signifies enrichment, i.e. more trinucleotide repeats than expected, and blue purification, i.e. less than expected, while the *P*-value is signified with color intensity with more bland color being less significant. White areas have no significant differences. See Additional file 3: Figure S2 for the complete analysis

activating proteins (GAPs) and guanine nucleotide-exchange factors (GEFs) by suppression (GAPs) or promotion (GEFs) of the GTPase' activity [75]. For instance, in humans, 81 GEFs and 67 GAPs [76] regulate the activity of the 22 Rho GTPases [77]. Some of the small GTPases are important for proper immune function [78, 79], by regulating chemotaxis and phagocytosis [80]. In mammals, the GTPase RhoA is important for TLR signaling, specifically for TLR2 and TLR4 [80]. Thus, between two populations of codfishes, adapted to different environments, there may potentially be variation in immune responses based on length variations of STRs in GEFs and GAPs.

Historic effective population size and STRs

Many marine fish with a pelagic life style are characterized by large effective population sizes [81]. Atlantic herring has an estimated effective population size of approximately 1 million and a nucleotide diversity of 0.32% [4], similar to cod with an effective population size around 400,000 and 0.39% nucleotide diversity and haddock at around 1.1 million and 0.54% nucleotide diversity (Table 1). Intriguingly, herring seems to have a high amount of STRs (Supplementary File E in [4]), suggesting that the life history strategies of cod, haddock and herring might facilitate a high density and frequency of STRs. The high effective population sizes in these species would imply low genetic drift and more efficient selection.

With around 760,000 STR loci in haddock and cod (Additional file 1: Table S2), the majority are likely to be highly polymorphic in such large haddock and cod populations. In a study of over 1000 human individuals, most of the 700,000 STR loci sequenced were polymorphic [31], although constraints were apparent for mutations in coding sequences [30]. Haddock and cod (Fig. 1) have at least ten times the historic effective population size of humans [55], and their high fecundity would generate many STR variants for each generation. We find trinucleotide indels in STRs in 2–3% of the genes, i.e., they have different length variants of the STRs in these genes. With such large effective populations and few barriers, genetic drift is weak, and local populations should respond to even weak selection [81]. There are studies suggesting STR loci are under selection in cod [82, 83]. Most tools for genome-wide investigations of selection have focused on SNPs, but methods for selection on STRs have been developed [84]. With high accuracy STR genotyping [85, 86] and resequencing data from different populations or controlled experiments over several generations, we suspect substantial numbers of STRs under selection will be found.

The *MHCI* and *TLR* repertoire in haddock and cod

In the first cod genome assembly, only two *MHCI* classical U-lineage genes were found, despite qPCR indicating around 100 copies [12]. Other investigations have also estimated a large number of *MHCI* copies in cod [11, 87, 88], but these have either investigated transcriptional data or read depth comparisons between *MHCI* loci and single-copy genes. Malmstørm et al. (2016) [11] estimated around 30 copies in haddock and 70 in cod. We found similar numbers to those predicted by [11] using our unitig assemblies of the same species; however in contrast a much lower number was found in the final assemblies. In the cod assembly, seven of the in total thirteen *MHCI* copies with complete alpha domains are located on unplaced contigs/scaffold in the gadMor2 assembly (data not shown). Their numbers are likely to be underestimated because the unplaced contigs/scaffold often have a higher read depth, indicating that these contain multiple, collapsed copies. Using PacBio reads in both the haddock and the cod assemblies likely substantially contributed to the more complete representation of *MHCI* genes, compared to the previous cod genome assembly. The Asian seabass, another assembly based on PacBio reads, resulted in "a more continuous cluster of MHC-class I genes compared to the well-assembled *G. aculeatus* [three-spined stickleback] genome" [26], highlighting the importance of long reads for properly capturing these regions of the genome. In contrast, the *TLR* repertoire is by and large similar between haddock and cod. The only main difference is found within *TLR22*; with twice as many copies in cod (10 vs. 5). We were unable to perform the domain-based search for *TLRs*, since they do not have a *TLR*-specific domain. The TIR domain (PFAM domain PF01582), the most likely candidate, is also found in the large interleukin-1 receptor family [89] and in adaptor proteins such as MyD88 and SARM [90].

The high copy number of *NLRs* in teleosts

In this study we enumerate genes (putative *NLRs*) with the NACHT (PFAM domain PF05729) and FISNA (PF14484) domains. These two domains together characterize a family of proteins substantially expanded in zebrafish with around 400 copies [60] and indications of substantial expansions in other teleosts as well [20, 91, 92].

For genome assemblers, identical or highly similar sequences occurring in multiple locations in a genome are indistinguishable from repeated sequence such as for example transposable elements. Depending on the sequencing strategy and assembler, these may introduce gaps into an assembly because the assembler is unable to place them correctly and they might be collapsed as a single contig/scaffold [23]. In general, genome assemblers might treat the large amount of *NLR* genes in these species as repeated sequence, and thus be unable to place them into scaffolds. For the species with genome assemblies in linkage groups or chromosomes, we looked at the contigs/scaffolds that were placed into these versus those that were not (Fig. 2). Even with the large number of genes (> 400), only 10% of the putative *NLRs* are unplaced for zebrafish. This is likely due to its sequencing and assembly strategy, with tiling of individually sequenced and assembled bacterial artificial chromosome clones [93]. For Atlantic cod, about 50% the contigs/scaffolds with putative *NLRs* are unplaced, and for stickleback about 15% are unplaced. The stickleback genome assembly is based on 9× coverage with Sanger sequencing reads [94], which may result in a more fragmented assembly than using PacBio reads (as for cod) or clones (zebrafish) because Sanger sequencing reads are shorter.

The numbers of putative *NLRs* from Fig. 2 should be interpreted with caution. It is likely that all species have some or several of the gene copies collapsed [20]. For Atlantic cod and haddock, we mapped reads back to the assembly, and investigated the coverage for all sequences (Fig. 3). There are many contigs/scaffolds with more than 5 times coverage compared to the average in the assemblies, and the numbers of putative *NLRs* are likely underestimated. Even though these two assemblies are highly contiguous and have been created with the use of PacBio reads, multi-copy genes such as *NLRs* may still be problematic. We also investigated the content of the unitig assemblies for Atlantic cod and haddock, and found similar numbers of *NLRs* between the two species (Table 5), however, many of these are likely pseudogenes due to stop codons. The difference between the unitig assemblies and the final assemblies are because of differences in assembly processes (Additional file 1: Note S1), where the final haddock assembly contains fewer short contigs. Most likely the *NLR* content of the two codfishes is highly similar. The numbers of *NLRs* are likely severely underestimated in most currently investigated ray-finned fish. Assemblies of higher quality are needed to properly investigate this intriguing family of innate immune genes.

It is unclear how such large gene families as the *NLRs* in zebrafish evolved [95]. In zebrafish, the majority of *NLRs* are located on one chromosome 4 arm [60] (Additional file 1: Table S6). Although the other assemblies are of lower quality than the zebrafish genome, there are no clear patterns of chromosomal enrichment in *NLRs* in other ray-finned fishes. Possible exceptions are medaka with 33 FISNA domains found on linkage group 2 (Additional file 1: Table S9) and stickleback with 12 FISNA and NACHT domains found on groupXIII (Additional file 1: Table S7). For Atlantic cod, the *NLRs* are evenly divided across linkage groups (Additional file 1: Table S5). Further, tetraodon (Additional file 1: Table S10) and spotted gar (Additional file 1: Table S8) have relatively few copies in total.

Conclusions

Our study provides new insight into elements of genomic architecture in two species of codfishes. The haddock genome contains an even higher density of STRs than the Atlantic cod genome. Further, certain classes of genes are enriched for STRs in both Atlantic cod and haddock, but not in the other published fish genome assemblies. With the large effective population sizes of cod and haddock, these STRs are likely polymorphic and represent a large reservoir of genetic variation. Additionally, for copy number estimations of highly expanded genes, such as the *NLR* genes, we discovered that the genome assemblies of most teleosts do not accurately represent these. Thus, the expanded nature of such gene families most likely confound genome assemblers, at least when based on Illumina reads or moderate coverage of PacBio reads. However, investigation of unitig assemblies of cod and haddock shows substantially higher copy numbers than the final assemblies. Most likely, the available teleost genome assemblies represent severe underestimations of the number of *NLR* genes. Better genome assemblies, i.e. created with sufficient long read coverage in combination with linked reads [96], optical mapping [64, 97] and/or chromosome conformation [25], should facilitate proper characterization of the *NLR* content as well as other teleost multi-copy genes, unraveling their evolutionary past.

Methods

Sampling and sequencing

The sequenced individual, a wild caught specimen approximately 1.3 kg belonging to the North-East Artic haddock population, was sampled near the Lofoten Islands (N68.04 E13.41), outside of its spawning season (in July 2009). We always aim to limit the effect of our research on populations and individuals. Whenever possible we collaborate with other sources, such as commercial fisheries or aquaculture farms, where samples can be harvested freely in combination with their normal business. This way, no animals need to be euthanized to serve our scientific purpose alone. The specimen used in this study comes from a wild population and was part of a larger haul of commercially fished individuals intended for human consumption. Following capture the fish was immediately stunned with a blunt object, then killed by bleeding, following standard procedure by local fishermen. Sampling in this manner does not fall under any specific legislation in Norway, but it is in accordance with the guidelines set by the 'Norwegian consensus platform for replacement, reduction and refinement of animal experiments' (www.norecopa.no). DNA was extracted from the spleen (stored on RNALater) using a standard high salt DNA extraction protocol.

200 bp insert size paired end libraries were constructed with Illumina DNA paired end sample preparation reagents and sequenced at the McGill University and Génome Québec Innovation Centre, both 100 bp long reads, with 322 M read pairs, 64 Gbp of sequence in total and 150 bp reads, with 224 M read pairs and 67 Gbp sequence. The 3 kbp (368 M read pairs, 74 Gbp) and 10 kbp (175 M read pairs, 35 Gbp) insert size libraries were prepared with the Illumina Mate Pair gDNA reagents and sequenced at the McGill University and Génome Québec Innovation Centre with 100 bp reads. All Illumina libraries were sequenced on the HiSeq 2000 using V3 chemistry.

PacBio SMRT sequencing was performed on a PacBio RS II instrument (Pacific Biosciences of California Inc., Menlo Park, CA, USA) at the Norwegian Sequencing Centre (NSC, www.sequencing.uio.no/). Long insert SMRTbell template libraries were prepared at NSC according to PacBio protocols. In total, 24 SMRT-cells were sequenced using P6v2 polymerase binding and C4 sequencing kits with 120 min acquisition. Approximately 16.4 Gbp of library bases were produced from 2.7 M reads with average read length of 5980 bp.

Assembly

Genome assembly

First, half the paired end library with read length 150 bp and insert size of 200 bp were used to satisfy the requirements of about 50× coverage in overlapping reads for ALLPATHS-LG [38]. Half the 200 bp insert size, 100 bp read length, was used as a jumping library, in addition to half the 3 kbp library and all of the 10 kbp library, again to approximate the requirements of the software. Release R48639 of ALLPATHS-LG was used.

Second, meryl from Celera Assembler 8.3rc2 [39] was used to count k-mers in the paired end Illumina libraries. All Illumina paired end reads were sequenced from the same DNA library, with insert size around 200 bp. Because of this overlapping reads were merged with FLASH v1.2.3 [98].

The merTrim program [28], also from Celera Assembler, was used to correct the output from FLASH, the merged and unmerged Illumina reads. The raw, uncorrected PacBio whole genome shotgun reads were separately trimmed by the overlap-based-trimming module in Celera Assembler [39]. The trimmed Illumina and PacBio reads were assembled together with Celera Assembler resulting in a contig assembly, following [28]. All Illumina reads were mapped to the contig assembly using BWA mem v0.7.9a [40], and the scaffold module from SGA (github snapshot June25th_2014) [41] was used to scaffold the contigs. All Illumina reads were again mapped to the scaffold assembly, and Pilon v1.16 [42] was applied, reducing some gaps and recalling consensus.

Transcriptome assembly

All RNA-seq data from Sørhus et al. (2017) [48] (Sequence Read Archive at NCBI with Accession ID:

PRJNA328092) was assembled with Trinity v2.0.6 [99]. This dataset consisted of paired 150 bp reads from RNA isolated from pools of embryos and larvae, and consisted of 1.6 G read pairs. Trinity was run with the '–trimmomatic' and '–normalize_reads' options, to remove adaptors and to normalize the coverage, respectively. No filtering was performed on the assembly before it was used in the MAKER annotation pipeline. The assembly consisted of 1,227,534 'genes', 1,766,998 transcripts with N50 at 960 bp. The total amount of assembled bases was 1.16 Gbp. All statistics are derived from the TrinityStats.pl script.

Validation of genome assembly.

CEGMA v2.4.010312 [43, 44] and BUSCO v2 [45] with an actinopterygii specific gene set were run on the genome assembly to asses the amount of conserved eukaryotic genes.

Annotation

Repeat library

A library of repeated elements was created as described in [28]. RepeatModeler v1.0.8, LTRharvest [100] part of genometools v1.5.7 and TransposonPSI were used in combination to create a set of putative repeats. Elements with only a match against an UniProtKB/SwissProt database and not against the database of known repeated elements included in RepeatMasker were removed. The remaining elements were classified and combined with known repeat elements from RepBase v20150807 [101].

Annotation

Three different ab initio gene predictors were trained. GeneMark-ES [102] v2.3e on the genome assembly, SNAP v20131129 [103] on the genes found by CEGMA, and AUGUSTUS v3.2.2 [104, 105] on the genes found by BUSCO. MAKER v2.31.8 [46, 47] used the trained gene predictors, the Trinity transcriptome assembly, the repeat library and proteins from UniProtKB/SwissProt r2016_3 [49] for a first pass [106] annotation of the genome assembly. The result of the first pass was used to retrain SNAP and AUGUSTUS, and a second iteration was performed using the same set-up.

The protein sequences from final output of MAKER was BLASTed against the UniProtKB/SwissProt proteins and InterProScan v5.4–47 [50] was used to classify protein domains in the protein sequences. Finally, the output of MAKER was filtered on AED, keeping only genes/proteins with an AED less than 0.5 (where 0.0 indicates perfect accordance between the gene model and evidence (mRNA and/or protein alignments), and 1.0 no accordance).

Finding orthologues

We downloaded all genome assemblies, cDNA and protein fasta files for all fishes at Ensembl release 81 (Amazon molly, cavefish, Atlantic cod (gadMor1), fugu, medaka, platyfish, spotted gar, stickleback, tetraodon, tilapia and zebrafish), and extracted the longest protein using a custom script (get_only_longest_protein_per_gene.py) because some annotations provide multiple proteins per gene. We did an all-against-all BLASTP of the protein sequences of all the Ensembl fishes in addition to the new cod and haddock annotated proteins, following the default options as set by OrthoFinder. The results of this were used as input to OrthoFinder v1.0.6 [52].

Investigating variants in the haddock and cod assemblies

Both haddock and cod were sequenced in the [11] study, and these 150 bp reads were mapped to the respective assemblies using BWA MEM v0.7.12 [40], and sorted using samtools v0.1.19 [107]. Bamtools v2.3.0 and the script 'coverage_to_regions.py' from FreeBayes v0.9.14 [53] were used to split the assembly into regions, and FreeBayes was run in parallel. Vcflib from a GitHub snapshot at 20140325 was used to filter the variants, and only variants with more than 20 in quality and 5 in depth were retained.

Estimating historic effective population size

A GitHub snapshot from August25th 2015 of PSMC [55] was used together with samtools v1.1 and bcftools v1.2 on the mapped reads, and historic effective population size was inferred for cod and haddock. The mutation rates were estimated along the branches of the phylogeny reported in [11] and the generation times were set to 10 years for cod and 6 years for haddock [56].

Identification of *TLRs*

Toll-like receptors (TLRs) are a key component of the innate immune response. The toll interleukine receptor (TIR) is the most conserved domain of the TLRs [108]. To determine candidate regions likely containing *TLR* genes, we aligned all TIRs protein sequences available on Ensembl and GenBank against the haddock and cod genome assemblies using TBLASTN from the BLAST+ suite [109] with an e-value cutoff of 1e-10. We then extracted 10,000 bp around the regions containing TIR like motifs. We used BLASTN to align coding sequences representative of all the *TLRs* classes against the candidate regions containing *TLR* copies. Here we report full-length *TLR* copies as well as partial copies (≥60% of the coding sequence).

Identification of *MHCI*

We used the alpha-3 domain of the MHCI complex to identify the candidate regions containing *MHCI* genes in both haddock and Atlantic cod. We used TBLASTN to align alpha-3 coding sequences from Atlantic cod and zebra fish (*Danio rerio*) against the haddock and Atlantic cod genome assemblies, with an e-value threshold of 1e-10. We then extracted the region located 10,000 bp around the putative alpha-3 domains. We used BLASTN to align the extracted regions against the non-redundant nucleotide database on NCBI. Regions containing the three alpha domains of MHCI (α1, α2 and α3) were used as a proxy to determine the number of *MHCI* gene copy number.

To better assess the differences between the unitig assemblies and the final assemblies, we translated the unitigs assemblies of melAeg and GM_CA454PB (both are basis for the final assemblies) into all six reading frames with transeq from Emboss v6.5.7 [57], and used the PFAM v31.0 [58] domain PF00129 (Class I Histocompatibility antigen, domains alpha 1 and 2; MHCI) in HMMER v3.1b2 [59] to search the unitig assemblies for putative *MHCI* genes.

Identification of *NLRs*

We ran InterProScan v5.4–47 [50] on the longest protein per gene to annotate protein domains. The default Ensembl annotation of these seemed out of date for several species, and with this procedure we had a more uniform dataset. We counted the occurrences of the PFAM v31.0 [58] domains PF05729 (NACHT domain) and PF14484 (Fish-specific NACHT associated domain, FISNA). In addition we translated the assemblies of all species into all six reading frames with transeq from Emboss v6.5.7 [57], and searched these with the NACHT and FISNA domains with HMMER v3.1b2 [59]. The species relationship in Fig. 2 is derived from [11] and we used ETE3 [110] to plot the dendogram.

We used v1.3.1 of samtools [107] with the 'depth –a –a' option to calculate the per base pair coverage of the assemblies, and used awk to calculate average depth per sequence and average for the whole assembly. We extracted all sequences with FISNA domains, and plotted length versus depth for these using ggplot2 [111] in the R environment.

As for *MHCI*, we searched the unitig assemblies of cod and haddock with the FISNA and NACHT domains.

STRs in the assemblies and coding regions

We used Phobos v3.3.12 [61] to detect all TRs with unit size 1–10 bp in the assemblies. The output was in Phobos native format that was processed with the sat-stat v1.3.12 program, yielding files with different statistics and a gff file. The other settings were as used in [28].

We counted the number of different STRs in genes and number of genes with STRs by using bedtools [112] and overlaps between STRs and genes. For cod and haddock, we also counted the number of overlaps between trinucleotide TRs, indels of size 3 and genes.

Enrichment of STRs in genes

For each gene ontology group we performed pairwise comparisons of the number of genes with STRs and total number of genes between the different species using Fisher's exact test (implemented in SciPy [113]). We corrected for multiple testing using the Benjamini-Yekutieli [114] procedure of False Discovery Rate as implemented in statsmodels (http://www.statsmodels.org/stable/index.html). Of 2748 gene ontology terms, we found significant differences in 74.

For each gene ontology group we also tested the enrichment or purification of STRs compared to amount of STRs all the genes in a species using goatools, and correcting for multiple testing with Benjamini-Yekutieli procedure of False Discovery Rate [62].

Additional files

Additional file 1: Note S1 and Table S1-S10. (PDF 595 kb)

Additional file 2: Figure S1. Pairwise Fisher's exact test between gene ontology (GO) groups and species. Significant differences were found in 74 of 2748 GO groups, i.e. one or more species had significantly higher proportion of genes with STRs in a GO group that other species as found by Fisher's exact test. In white and light blue areas there are no significant differences, but in dark blue areas there are significant differences between two species. (PDF 244 kb)

Additional file 3: Figure S2. The terms that are significantly enriched for genes with trinucleotide tandem repeats in different species, those repeats that can vary in length without causing frameshifts in the protein. Only tests with $P < 0.01$ are colored. Red signifies enrichment, more trinucleotide repeats than expected, and blue purification, less than expected. The *P*-value is signified with color intensity with more bland color being less significant. White areas have no significant differences. (PDF 39 kb)

Acknowledgements

All computational work was performed on the Abel Supercomputing Cluster (Norwegian metacenter for High Performance Computing (NOTUR) and the University of Oslo) operated by the Research Computing Services group at USIT, the University of Oslo IT-department (http://www.hpc.uio.no/). Sequencing library creation and high throughput sequencing was carried out at the Norwegian Sequencing Centre (NSC), University of Oslo, Norway. We especially thank Marianne H. S. Hansen for DNA extraction and Ave Tooming-Klunderud for PacBio RS II library preparation and sequencing, both affiliated NSC, University of Oslo. We also thank Mark Ravinet for a critical reading of the manuscript.

Funding

This research was supported by the Norwegian Research Council under the projects "Functional and comparative immunology of a teleosts world without MHCII (#222378/F20)" led by Prof. Kjetill S. Jakobsen (University of Oslo) and "Assessment of long-term effects of oil exposure on early life stages of Atlantic haddock using state-of -the art genomics tools in combination with fitness observations" (#234367/E40) led by Dr. Sonnich Meier (IMR). The funding body had no part in the design of the study, collection, analysis and interpretation of data nor in writing the manuscript.

Availability of data and materials

The genome assembly and annotation are available from FigShare: https://doi.org/10.6084/m9.figshare.5182861

Illumina sequencing reads are available from ENA at http://www.ebi.ac.uk/ena/data/view/PRJEB21701.

The genome assembly is also available from ENA at https://www.ebi.ac.uk/ena/data/view/PRJEB25074.

Authors' contributions

OKT created the genome assembly and annotated it, performed all analyses involving STRs, the analyses using protein domains and wrote the first draft of the manuscript. MSOB performed the BLAST-based analyses with assistance of MHS. MHS and RBE performed preliminary analysis of NOD-like receptor genes. ES sampled and extracted RNA for RNA-seq. OKT, AJN and SJ designed the sequencing strategy. KSJ and SJ oversaw the project. SM, RBE and SJ conceived the study. All authors read and approved the final manuscript.

Ethics approval

We have adhered to all local, national and international regulations and conventions, and we respected normal scientific ethical practices. The specimen used in this study comes from a wild population and was part of a larger haul of commercially fished individuals intended for human consumption. Following capture the fish was immediately stunned with a blunt object, then killed by bleeding, following standard procedure by local fishermen. Sampling in this manner does not fall under any specific legislation in Norway, but it is in accordance with the guidelines set by the 'Norwegian consensus platform for replacement, reduction and refinement of animal experiments' (www.norecopa.no).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway. [2]Institute of Marine Research, Bergen, Norway. [3]Biomedical Informatics Research Group, Department of Informatics, University of Oslo, Oslo, Norway.

Received: 16 August 2017 Accepted: 22 March 2018
Published online: 10 April 2018

References

1. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29:51–63.
2. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513:375–81.
3. Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Comms. 2014;5:5770.
4. Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, et al. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. elife. 2016;5:311.
5. Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, et al. The seahorse genome and the evolution of its specialized morphology. Nature. 2016;540:395–9.
6. Small CM, Bassham S, Catchen J, Amores A, Fuiten AM, Brown RS, et al. The genome of the Gulf pipefish enables understanding of evolutionary innovations. Genome Biol. 2016;17:258.
7. Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, et al. The African coelacanth genome provides insights into tetrapod evolution. Nature. 2013;496:311–6.
8. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48:427–37.
9. Olsen E, Aanes S, Mehl S, Holst JC, Aglen A, Gjosaeter H. Cod, haddock, saithe, herring, and capelin in the Barents Sea and adjacent waters: a review of the biological value of the area. ICES J Mar Sci. 2010;67:87–101.
10. FAO. The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all. Rome. 2016;1–204.
11. Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, et al. Evolution of the immune system influences speciation rates in teleost fishes. Nat Genet. 2016;48:1204–10.
12. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011;477:207–10.
13. Solbakken MH, Rise ML, Jakobsen KS, Jentoft S. Successive losses of central immune genes characterize the Gadiformes' alternate immunity. Genome Biol Evol. 2016;8:3508–15.
14. O'Neill LAJ, Golenbock D, Bowie AG. The history of toll-like receptors — redefining innate immunity. Nat Rev Immunol. 2013;13:453–60.
15. Solbakken MH, Tørresen OK, Nederbragt AJ, Seppola M, Gregers TF, Jakobsen KS, et al. Evolutionary redesign of the Atlantic cod (*Gadus morhua* L.) toll-like receptor repertoire by gene losses and expansions. Sci Rep. 2016;6:25211.
16. Solbakken MH, Voje KL, Jakobsen KS, Jentoft S. Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. Proc Biol Sci. 2017;284:20162810.
17. Malmstrøm M, Jentoft S, Gregers TF, Jakobsen KS. Unraveling the evolution of the Atlantic cod's (*Gadus morhua* L.) alternative immune strategy. PLoS One. 2013;8:e74004.
18. Motta V, Soares F, Sun T, Philpott DJ. NOD-like receptors: versatile cytosolic sentinels. Physiol Rev. 2015;95:149–78.
19. Bonardi V, Cherkis K, Nishimura MT, Dangl JL. A new eye on NLR proteins: focused on clarity or diffused by complexity? Curr Opin Immunol. 2012;24:41–50.
20. Stein C, Caccamo M, Laird G, Leptin M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. Genome Biol. 2007;8:R251.
21. Lange C, Hemmrich G, Klostermeier UC, López-Quintero JA, Miller DJ, Rahn T, et al. Defining the origins of the NOD-like receptor system at the base of animal evolution. Mol Biol Evol. 2011;28:1687–702.
22. Rast JP, Smith LC, Loza-Coll M, Hibino T, Litman GW. Genomic insights into the immune system of the sea urchin. Science. 2006;314:952–6.
23. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Rev Genet. 2012;13:36–46.
24. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8:61–5.
25. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.
26. Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, van Heusden P, et al. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. PLoS Genet. 2016;12: e1005954. Richardson PM, editor
27. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A new chicken genome assembly provides insight into avian genome structure. G3. 2016;7:109–17.
28. Tørresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, et al. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics. 2017;18:95.
29. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature Rev Genet. 2004;5:435–45.
30. Gymrek M, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet. 2017;49:1495–501.
31. Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014;24:1894–904.
32. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016;48:22–9.
33. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445–77.
34. Mularoni L, Ledda A, Toll-Riera M, Albà MM. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res. 2010;20:745–54.
35. Albà MM, Santibáñez-Koref MF, Hancock JM. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol. 1999;49: 789–97.
36. Huntley MA, Clark AG. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. Mol Biol Evol. 2007;24:2598–609.
37. Zhao Z, Guo C, Sutharzan S, Li P, Echt CS, Zhang J, et al. Genome-wide analysis of tandem repeats in plants and green algae. G3. 2014;4:67–78.

38. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2011;108:1513–8.
39. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24:2818–24.
40. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 [q-bio.GN]. 2013.
41. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22:549–56.
42. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
43. Parra G, Bradnam KR, Ning Z, Keane T, Korf IF. Assessing the gene space in draft genomes. Nucleic Acids Res. 2009;37:289–97.
44. Parra G, Bradnam KR, Korf IF. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.
45. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
46. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
47. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol American Society of Plant Biologists. 2014;164:513–24.
48. Sørhus E, Incardona JP, Furmanek T, Goetz GW, Scholz NL, Meier S, et al. Novel adverse outcome pathways revealed by chemical genetics in a developing marine fish. elife. 2017;6:e20707.
49. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–12.
50. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–40.
51. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 2009;10:67.
52. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
53. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN]. 2012.
54. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Rev Genet. 2009;10:195–205.
55. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.
56. Durant JM, Hjermann DØ. Age-structure, harvesting and climate effects on population growth of Arcto-boreal fish stocks. Mar Ecol Prog Ser. 2017;577: 177–88.
57. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–77.
58. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
59. Eddy SR. Accelerated profile HMM searches. PLoS Comp Biol. 2011;7: e1002195.
60. Howe K, Schiffer PH, Zielinski J, Wiehe T, Laird GK, Marioni JC, et al. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. Open Biol. 2016;6:160009–224.
61. Mayer C, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in *Daphnia pulex* - a comparative approach. BMC Genomics. 2010;11:277.
62. Tang H, Klopfenstein D, Pedersen B, Flick P, Sato K, Ramirez F, et al. GOATOOLS: tools for gene ontology. Zenodo. 2015. https://doi.org/10.5281/zenodo.31628.
63. Li J, Bian C, Hu Y, Mu X, Shen X, Ravi V, et al. A chromosome-level genome assembly of the Asian arowana, *Scleropages formosus*. Sci Data. 2016;3:160105.
64. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538:243–47.
65. Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotech. 2018. Advanced online publication. https://www.nature.com/articles/nbt.4060.
66. Adams RH, Blackmon H, Reyes-Velasco J, Schield DR, Card DC, Andrew AL, et al. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. Genome. 2016;59:295–310.
67. Jiang Q, Li Q, Yu H, Kong L. Genome-wide analysis of simple sequence repeats in marine animals—a comparative approach. Mar Biotechnol. 2014; 16:604–19.
68. Star B, Hansen MH, Skage M, Bradbury IR, Godiksen JA, Kjesbu OS, et al. Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. Sci Technol Archaeol Res. 2016;2:36–45.
69. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447: 932–40.
70. Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC. Origin, evolution and genome distribution of microsatellites. Genet Mol Biol. 2006;29:294–307.
71. Legendre M, Pochet N, Pak T, Verstrepen KJ. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 2007;17:1787–96.
72. Gemayel R, Chavali S, Pougach K, Legendre M, Zhu B, Boeynaems S, et al. Variable glutamine-rich repeats modulate transcription factor activity. Mol Cell. 2015;59:615–27.
73. Takai Y, Sasaki T, Matozaki T. Small GTP-Binding Proteins. Physiol Rev. 2001; 81:153–208.
74. van Dam TJP, Bos J, Snel B. Evolution of the Ras-like small GTPases and their regulators. Small GTPases. 2014;2:4–16.
75. Rossman KL, Der CJ, Sondek J. GEF means go: turning on RHO GTPases with guanine nucleotide-exchange factors. Nat Rev Mol Cell Biol. 2005;6:167–80.
76. Zaritsky A, Tseng Y-Y, Rabadán MA, Krishna S, Overholtzer M, Danuser G, et al. Diverse roles of guanine nucleotide exchange factors in regulating collective cell migration. J Cell Biol. 2017; jcb.201609095
77. Ridley AJ. Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. Trends Cell Biol. 2006;16:522–29.
78. Johnson DS, Chen YH. Ras family of small GTPases in immunity and inflammation. Curr Opin Pharmacol. 2012;12:458–63.
79. Scheele JS, Marks RE, Boss GR. Signaling by small GTPases in the immune system. Immunol Rev. 2007;218:92–101.
80. Bokoch GM. Regulation of innate immunity by rho GTPases. Trends Cell Biol. 2005;15:163–71.
81. Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D. Population genomics of marine fishes: identifying adaptive variation in space and time. Mol Ecol. 2009;18:3128–50.
82. Nielsen EE, Hansen MM, Meldrup D. Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in nonmodel organisms. Mol Ecol. 2006;15:3219–29.
83. Eiríksson GM, Árnason E. Spatial and temporal microsatellite variation in spawning Atlantic cod, *Gadus morhua*, around Iceland. Can J Fish Aquat Sci. 2013;70:1151–8.
84. Haasl RJ, Payseur BA. Microsatellites as targets of natural selection. Mol Biol Evol. 2012;30:mss247–98.
85. Kristmundsdóttir S, Sigurpálsdóttir BD, Kehr B, Halldorsson BV. popSTR: population-scale detection of STR variants. Bioinformatics. 2016:btw568.
86. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;39:1.
87. Persson A-C, Stet RJM, Pilström L. Characterization of MHC class I and *β*2-microglobulin sequences in Atlantic cod reveals an unusually high number of expressed class I genes. Immunogenetics. 1999;50:49–59.
88. Miller KM, Kaukinen KH, Schulze AD. Expansion and contraction of major histocompatibility complex genes: a teleostean example. Immunogenetics. 2001;53:941–63.
89. Ve T, Williams SJ, Kobe B. Structure and function of toll/interleukin-1 receptor/resistance protein (TIR) domains. Apoptosis. 2014;20:250–61.
90. O'Neill LAJ, Bowie AG. The family of five: TIR-domain-containing adaptors in toll-like receptor signalling. Nat Rev Immunol. 2007;7:353–64.
91. Xu T, Xu G, Che R, Wang R, Wang Y, Li J, et al. The genome of the miiuy croaker reveals well-developed innate immune and sensory systems. Sci Rep. 2016;6:21902.
92. Laing KJ, Purcell MK, Winton JR, Hansen JD. A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. BMC Evol Biol. 2008;8:42.

93. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503.
94. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61.
95. Schiffer PH, Gravemeyer J, Rauscher M, Wiehe T. Ultra large gene families: a matter of adaptation or genomic parasites? Life. 2016;6:32.
96. Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 2017;24:2041.
97. Howe K, Wood JM. Using optical mapping data for the improvement of vertebrate genome assemblies. GigaScience. 2015;4:10.
98. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63.
99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
100. Ellinghaus D, Kurtz S, Willhoeft U. *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:1.
101. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
102. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33:6494–506.
103. Korf IF. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
104. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19:ii215–25.
105. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
106. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 2014;48:4.11.1–4.11.39.
107. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
108. Mikami T, Miyashita H, Takatsuka S, Kuroki Y, Matsushima N. Molecular evolution of vertebrate toll-like receptors: evolutionary rate difference between their leucine-rich repeats and their TIR domains. Gene. 2012;503:235–43.
109. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
110. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635–8.
111. Wickham H. ggplot2: elegant graphics for data analysis. 2016. New York: Springer-Verlag; 2016.
112. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
113. Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001. http://www.scipy.org. Accessed 7 July 2017.
114. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29:1165–88.