RESEARCH

An improved genome assembly uncovers a prolific tandem repeat structure in Atlantic cod

Ole K. Tørresen[1], Bastiaan Star[1], Sissel Jentoft[1,2], William Brynildsen Reinar[1], Harald Grove[3], Jason R. Miller[4], Brian P. Walenz[5], James Knight[6], Jenny M. Ekholm[7], Paul Peluso[7], Rolf B. Edvardsen[8], Ave Tooming-Klunderud[1], Morten Skage[1], Sigbjørn Lien[3], Kjetill S. Jakobsen[1] and Alexander J. Nederbragt[1,9*]

Abstract

Background: The first Atlantic cod (*Gadus morhua*) genome assembly published in 2011 was one of the early genome assemblies exclusively based on high-throughput 454 pyrosequencing. Since then, rapid advances in sequencing technologies have led to a multitude of assemblies generated from complex genomes, although many of these are of a fragmented nature with a significant fraction of bases in gaps. The development of long-read sequencing and improved software enable the generation of more contiguous genome assemblies.

Results: By combining data from Illumina, 454 and the longer PacBio sequencing technologies, as well as integrating the results of multiple assembly programs, we have created a substantially improved version of the Atlantic cod genome assembly. The sequence contiguity of this assembly has increased fifty-fold and the proportion of gap-bases has been reduced 15-fold. Compared to other vertebrates, the assembly contains an unusual high density of tandem repeats (TRs). Indeed, retrospective analyses reveal that gaps in the first genome assembly were largely associated with these TRs. We show that 21 % of the TRs across the assembly, 19 % in the promoter regions and 12 % in the coding sequences are heterozygous in the sequenced individual.

Conclusions: The use of multiple assembly programs combined with inclusion of PacBio reads drastically improved the Atlantic cod genome assembly by successfully resolving long TRs. The high frequency of heterozygous TRs within or in the vicinity of genes in the genome indicate a considerable standing genomic variation in Atlantic cod populations, which likely is of evolutionary importance.

Keywords:
assembly algorithms; PacBio; long-read sequencing technology; microsatellites; repetitive DNA; dinucleotide repeats; assembly consolidation; assembly assessment; heterozygosity; indel polymorphism

Background

The speed and affordability of sequencing and improved software, including more efficient genome assemblers, have lead to a democratization of genomics, enabling individual research groups to create *de novo* genome assemblies [1]. The first published *de novo* assemblies for non-model organisms using pure massively parallel sequencing approaches (Illumina and 454) appeared in 2010-2011 and include diverse species such as giant panda [2], turkey [3], woodland strawberry [4] and Atlantic cod [5]. Numerous genome assemblies from a myriad of non-model plants, invertebrates and vertebrates are now available, including examples of genomes that are difficult to assemble, e.g. the extremely large genomes of bread wheat [6] and Norway spruce [7], the highly heterozygous genome of oyster [8] and the tetraploid and repetitive salmon genome [9]. These genome assemblies have provided new insights into a range of fundamental biological questions, including the first example of a vertebrate immune system, that of Atlantic cod, without MHC (major histocompatibility complex) class II [5], the untangling of the events of multiple hybridizations shaping the ancestral genomes of bread wheat prior to domestication [10] and the multiple avian genomes resolving bird phylogeny, their radiations and investigation of the genetic basis of complex traits [11, 12]. Despite the tremendous impacts of the high throughput sequencing generated genomes, many of these assemblies are

[*] Correspondence: lex.nederbragt@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

of varying completeness, depending on the purpose for which they have been obtained [13, 14]. In the examples given above, the sizes of the scaffold sequences are usually far shorter than chromosome arm lengths. Most of these genomes have scaffold N50 lengths (i.e., half the assembly is in scaffolds of this length or longer) in the range of 400 kbp – 1.5 Mbp, although some avian genomes have N50 scaffold lengths up to 10 Mbp, approaching chromosome arm lengths. However, contig N50 lengths are far smaller and in the range of 3 kbp - 55 kbp.

The presence of repetitive DNA is the most important factor that leads to fragmented genome assemblies [14, 15]. Assembly algorithms struggle to resolve repetitive regions if these are longer than the read length, and this problem particularly affects the assembly of sequencing data from short-read technologies such as the Illumina platform [14–16]. Repetitive regions can be divided into two classes, interspersed and tandem repeats. Interspersed repeats, including transposable elements (TEs), occur across the genome and are present in all vertebrate genomes, comprising from 5 % to 55 % of their assemblies [17]. Tandem repeats (TRs) are sequences with a repeat unit repeated more than two times in tandem. TRs typically occupy 0.5 to 3 % of eukaryotic genomes, and can be classified into microsatellites, simple repeats, or short tandem repeats (STRs, 1-6 bp, or 1-9 bp repeat unit size); minisatellites (10-100 bp) and satellite repeats (>100 bp repeat unit size) [18]. TRs mutate by adding or removing full repeat units and their mutation rates, at least for STRs, are 10 to 10,000 fold higher compared to the average rates in the genome [19]. The heterozygosity caused by TR mutations, in addition to other types of heterozygosity, are also likely to have confounding effects on the contiguity of genome assemblies [14, 15].

Long-read sequencing technologies such as PacBio and Oxford Nanopore address the drawbacks of short-read technologies by being able to read through larger repeat regions, and are therefore particularly well-suited for *de novo* genome assembly [14, 20]. Including moderate amounts of PacBio coverage (5-20x) can dramatically improve the contiguity of an assembly in combination with other sequencing data [21, 22]. More extensive coverage (>50x) has enabled assemblies of vertebrate genomes reconstructing nearly complete chromosome arms [23–25], although the associated costs may be economically prohibitive. Nonetheless, a sequencing strategy including long-reads is recommended to aid reducing the fragmentation typical of *de novo* genome assemblies based on a short-read technology only. Regardless of sequencing strategy, usage of a genetic linkage map, or an optical map, can place contigs or scaffolds into chromosome-sized reconstructions, called linkage groups, a prerequisite for large-scale genome comparisons between species [26].

The first release of the Atlantic cod (*Gadus morhua*) genome was sequenced and assembled with the 454 sequencing technology only [5] and annotated by the Ensembl Project [27] (gadMor1). The 832 Mbp assembly was fragmented, with a contig N50 2.3 kbp and 27 % of bases in gaps. The genome contained 17.8 % transposable elements and 5.9 % tandem repeats (Supplementary Table 6 in [5]). An increased abundance of TRs (unit size 1-4 bp) at the contig termini (32 %), and at the gaps in scaffolds (24 %, Supplementary Note 7 in [5]) indicated these repeats contributed to the observed level of fragmentation.

A more contiguous reference genome for Atlantic cod, preferably with chromosome-level reconstructions, will facilitate re-sequencing efforts addressing population genomics investigations, including detecting structural variants, introgression and hybridization between species, as well as improve comparative genomic investigations relying on synteny. Moreover, it would also enable an annotation with more complete gene models and allow for a better understanding of the lack of sequence contiguity in gadMor1. To achieve this, we created several assemblies using different combinations of Illumina, 454 and PacBio sequencing technologies as well as Sanger BAC-end sequences, using a suite of assembly programs. As often is the case [28–30], no single assembly outperformed the others in various criteria (N50 contig/scaffold length, gene content, agreement with a genetic linkage map, accordance with read data), and thus a reconciled assembly was created to integrate the best characteristics of four draft assemblies. This new assembly, denoted gadMor2, has a fifty-fold improvement of the contig N50 length of gadMor1, and eight times as long scaffold N50, with 1.7 % bases in gaps compared to 27 % in gadMor1. A linkage map (personal communication, Sigbjørn Lien) was used to order and orient the scaffolds into linkage groups. The new genome assembly and annotation reveal a high content of tandem repeats compared to other vertebrates, across the genome, and most notably in promoter regions and within amino acid coding sequences. Many of these TRs are heterozygous and we propose that this has implications for understanding local adaptation at a population level.

Results

An improved genome assembly for Atlantic cod

In addition to already existing sequencing data for the wild-caught individual from the North East-Arctic population described in [5] (~40x Roche/454 and ~0.1x Sanger BAC-ends), we added sequencing data from

Table 1 Overview of assembly statistics.

Assembly	Total size assembly (Mbp)	N50 contig (kbp)	N50 scaffold (Mbp)	Percentage gap bases	CEGMA[1]	BUSCO[2]	REAPR[3]	FRC^{bam}[4]	Potential conflict (sequences)[5]
gadMor1[6]	832	2.3	0.14	26.9	444 (96.9 %)	3,308 (89.4 %)	2,547	4,210,772	76
ALPILM	660	4.4	0.16	28.7	424 (92.6 %)	3,016 (81.6 %)	19,787	2,182,096	122
NEWB454	656	6.2	1.30	24.4	435 (95.0 %)	3,109 (84.1 %)	18,117	2,044,008	26
CA454ILM	647	9.9	0.50	3.49	447 (97.5 %)	3,379 (91.4 %)	7,406	1,351,500	96
CA454PB	682	95	0.27	1.62	431 (94.1 %)	3,310 (89.5 %)	8,617	1,508,054	188
gadMor2[7]	643	116	1.15	1.69	435 (95.0 %)	3,447 (93.2 %)	7,359	1,248,792	15

[1] CEGMA annotates 458 highly conserved eukaryotic genes
[2] BUSCO annotates 3,698 actinopterygii specific genes
[3] REAPR analyses the discordance between the expected order, orientation and distance of mapped paired reads, with detected potential errors, fewer is better
[4] FRC^{bam} uses a similar approach as REAPR, with total number of features (i.e., potential assembly problems), fewer is better
[5] number of sequences mapping to more than one linkage group or to multiple linkage groups, fewer is better
[6] from [5]
[7] 93 % of the gadMor2 assembly is additionally oriented and ordered into 23 linkage groups (Supplementary Table 3)

Illumina (~480x coverage) and PacBio (~19x coverage) (Supplementary Table 1) obtained from DNA isolated from the same individual. Different assembly strategies were used: a Newbler assembly with 454 and Sanger BAC-end sequences as input (referred to as NEWB454 for short), an ALLPATHS-LG [31] assembly with the Illumina sequences only (ALPILM), a Celera Assembler [32] assembly with 454 and Illumina sequences (CA454ILM) and a Celera Assembler assembly with 454 paired reads, Illumina reads and raw, uncorrected PacBio reads (CA454PB) (Supplementary Table 1). For each of the individual assemblies, different combinations of the assembly improvement programs Pilon [33] and PBJelly [34] were applied to improve the consensus sequence and to close gaps (Supplementary Table 2). The properties of these assemblies were assessed using multiple tools: those based on the mapping of read datasets to an assembly: FRC^{bam} [35], REAPR [36]; comparing a transcriptome to an assembly: Isoblat (using the Newbler transcriptome, see Methods) [37]; comparing the assembly to a linkage map (see Methods); and determining presence and completeness of conserved eukaryotic and Actinopterygii (ray-finned fishes) gene sets: CEGMA [38] and BUSCO [39] (Supplementary Table 2).

Based on these evaluations, each separate assembly had distinct properties, and none clearly outperformed any of the other on all metrics. For instance, the NEWB454 assembly has the largest scaffold N50 and the lowest number of conflict sequences (Figure 1, Table 1). In contrast, the CA454PB outperforms the other assemblies based on contig N50, yet has a lower scaffold N50 and higher number of sequences conflicting with the linkage map (sequences that map to two linkage groups) (Table 1). Existing assembly reconciliation tools are limited to combining just two assemblies [40, 41] and did not perform satisfactory. To obtain the best possible assembly, i.e., to integrate the information recovered by the different assemblies, we developed a novel assembly reconciliation method. This method involved an all against all alignment of the assemblies using Mugsy [42] after splitting the different assemblies when in conflict with the linkage map (see Methods) and removing sequences shorter than 1000 bp. The resulting alignment graph structure was traversed following the path from one of the original assemblies (CA454ILM, the one with the most genes found with CEGMA and BUSCO), outputting the sequence from the assembly with the least gaps (CA454PB), while using the alignments with ALPILM and NEWB454 in the graph to close gaps and extend

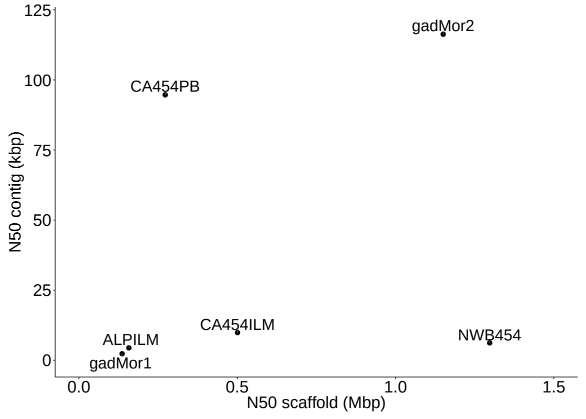

Figure 1 Contig and scaffold N50 of the different cod assemblies. gadMor2 was created by following the sequences in CA454ILM in a path through a graph consisting of multiple alignments, and outputting the contig sequences from CA454PB for each alignment. NEWB454 and ALPILM were used to extend the scaffolds (Table 1).

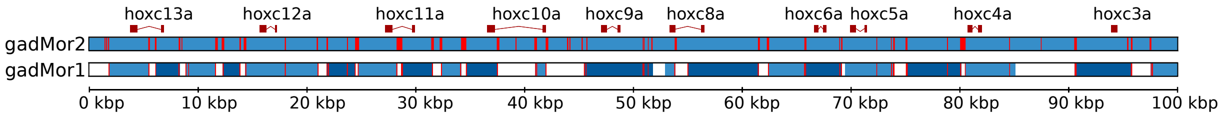

Figure 2 The HoxC cluster in gadMor1 and gadMor2. Blocks of dark and light blue are contig sequences, white blocks are gaps and red lines are tandem repeats. Gene models are sketched at the top of the figure. This region is a single contig in gadMor2 and 21 contigs in gadMor1. Tandem repeats are at the borders between almost all gaps and contigs in gadMor1.

scaffolds. The scaffold module from SGA [43] was applied on the resulting merged assembly using all paired reads (Illumina, 454 and sequenced BAC-ends), Pilon [33] was used to improve per-base accuracy and to close or reduce gaps. The resulting assembly was ordered and oriented based on a linkage map of 9355 SNPs (personal communication, Sigbjørn Lien) placing 93 % of the sequences into 23 linkage groups (Supplementary Table 3). We call this assembly gadMor2. Comparisons of assembly statistics for the final, reconciled assembly (gadMor2) and the original four (CA454ILM, CA454PB, ALPILM and NEWB454), show that gadMor2 outperforms all other assemblies on all quality metrics apart from scaffold N50 (ranked 2^{nd}) and CEGMA gene content (ranked 3^{rd}, Table 1, Figure 1). Based on an overall assessment of quality, gadMor2 combines the best features of each of the four original assemblies without loss of quality (Table 1).

The gadMor2 assembly has a fifty-fold larger contig N50 and eight-fold larger scaffold N50 compared to the gadMor1 assembly [5]. This has dramatic consequences for the sequence contiguity; for instance, a 100kbp region containing the HoxC cluster is a single contig in gadMor2, while it previously consisted of 21 contigs and 20 gaps in gadMor1 (Figure 2).

Genome size

Estimation of genome size with odd-sized kmers from 17 to 31 with SGA PreQC [44] on the 300 bp insert size, 100 bp reads, paired end Illumina reads (about 150x coverage), resulted in a genome estimate of 613 Mbp±11 Mbp (Supplementary Table 4), while the assembler ALLPATHS-LG estimated the genome to be 651 Mbp based on the k-mer distribution of the 180 bp insert size, 100 bp reads, paired end Illumina reads (about 52x coverage). Both estimates are lower than previous ones based on Feulgen Image Analysis Densitometry at 0.93 pg or 910 Mbp [45, 46] and a k-mer analysis based on 454 reads, which resulted in a 830 Mbp estimate [5]. Although the assembly size of the gadMor1 Atlantic cod genome assembly at Ensembl is 832 Mbp with 26.9 % gaps [5], the amount of sequence in contigs for this assembly is 608 Mbp, considerably closer to the SGA PreQC estimate. The likely explanation for the large size of gadMor1 is that many of the contigs could not be placed into a scaffold, and a gap was created at that locus instead. These unplaced contigs are included in the output contigs, resulting in loci represented twice in the assembly, once as a gap and once as a contig. The assemblies created in this study all span approximately 650 Mbp, which is similar to the ALLPATHS-LG estimation. Given the consistently lower range of values of the current k-mer analyses and genome assembly lengths, we here choose 650 Mbp as a likely more accurate Atlantic cod genome size estimate.

Annotation

We annotated 83,505 gene models with MAKER2 [47, 48], obtaining a final set of 23,243 predicted genes after discarding gene models with low support (see Methods). Compared to gadMor1 (20,095 predictions) [5], the gadMor2 annotation contains more predicted genes and significantly more sequence in the predicted transcriptome (32.2 Mbp and 52.9 Mbp, respectively). The predicted transcripts are substantially longer and without any gaps (Table 2). A genome browser enabling access to the genome and the annotation is available [49].

Heterozygosity

Illumina paired-end reads with 300 bp insert size and 100 bp read length were mapped to the gadMor2 assembly using BWA-MEM [54], and 2,621,997 SNPs (single nucleotide polymorphisms), 90,292 MNPs (multiple nucleotide polymorphisms), 631,063 indels (insertions and deletions) and 169,181 complex regions (composite insertion and substitution events) with quality 20 or better were called using FreeBayes [55]. With 2,621,997 SNPs, this corresponds to a (SNP) heterozygosity rate of 4.07×10^{-3} (one segregating site every 246 bp). The indel rate in Atlantic cod is at 0.98×10^{-3} (one indel every 1020 bp on average, Table 3).

We also called indels based on PacBio sequencing reads using blasr [56] and PBHoney [57]. 70,278 indels of size 20 bp or larger were found, at a rate of 0.1×10^{-3} indels/base, or one indel of size 20 bp or larger every 10000 bp on average.

Table 2 Comparison between the gene annotations of gadMor1 and gadMor2.

Assembly	Total size transcriptome (Mbp)[1]	Number of genes	N50 length (bp)[2]	Amount gap bases (Mbp)[3]	BUSCO [4]
gadMor1	32.2 (24.8)	22,618[5]	1,854 (1,398)	1.7	2,947 (79.7 %)
gadMor2	52.9 (33.4)	23,246[6]	3,239 (1,995)	0	2,714 (73.4 %)

[1] sum of bases in transcripts with UTRs (without UTRs)
[2] half the transcriptome in sequences of this length or longer, with UTRs (without UTRs)
[3] gaps represented as 'N's in annotated transcripts
[4] number (percentage) of conserved actinopterygii genes detected out of a total of 3,698
[5] when excluding pseudogenes, alternative transcripts, etc., the number of protein-coding genes is 20,095
[6] protein-coding genes only

Table 3 Comparison of the SNP and indel rates of selected organisms.

Species	SNP rate (SNPs/base)	Indel rate (indels/base)	N50 contig (kbp)	N50 scaffold (Mbp)
gadMor2	4.07×10^{-3}	0.98×10^{-3}	116	1.15
Stickleback[1]	1.43×10^{-3}	NA	83.2	10.8
Miiuy croaker[2]	2.24×10^{-3}	0.61×10^{-3}	73.3	1.15
Atlantic herring[3]	3.2×10^{-3}	NA	21.3	1.84
Ciona savignyi[4]	46×10^{-3}	NA	12	0.192
Ciona savignyi[5]	46×10^{-3}	NA	47	0.989

[1] from [50]
[2] from [51]
[3] from [52]
[4] from [53]
[5] from [53], with haplotype assembly and merging

Repeat content

We created a repeat library using a combination of RepeatModeler [58], LTRharvest [59], LTRdigest [60] and TransposonPSI [61] and known eukaryotic transposable elements sequences from RepBase [62] (see Methods). This library masked 31.28 % of the genome assembly (Table 4), with 22.86 % of the genome classified as interspersed repeats (most often TEs) and 7.96 % as TRs (used here as dinucletide to hexanucleotide repeats, at least 20 bp long), both classifications higher than for gadMor1 (17.8 % and 5.9 % respectively, Supplementary Table 6 in [5]), indicating a more complete genome assembly.

Tandem repeat content

We investigated to what extent different assemblers and sequencing technologies resulted in a difference in annotated TRs. Phobos [18] was used to find all TRs with a unit size of 1-50 bp, at least 13 bp long (different from the TRs classified above), in the different cod assemblies (Figure 3 and Table 5). Assemblies created with the Celera Assembler have the largest amount of TRs (Figure 3).

The most prominent class of TRs in gadMor2 is dinucleotide TRs, which make up 48.7 % of all annotated repeats, followed by mononucleotide, trinucleotide and tetranucleotide repeats comprising only 7.6 %, 6.3 % and 6.3 %, respectively (Figure 4). The average length of dinucleotide repeats is 84.4±87.2 bp, at an average 97.3 % identity. In total, dinucleotide repeats make up 5.7 % of the entire gadMor2 assembly. NEWB454 and ALPILM have significantly lower amount of, and shorter, TRs annotated than the two assemblies created with Celera Assembler, CA454ILM and CA454PB (Table 5).

A comparative analysis of gadMor2 to all genomes in Ensembl (release 81, excluding gadMor1), including the genome of California sea hare (which contains a large amount of TRs [63]), shows that the Atlantic cod genome assembly has approximately density of TRs three-fold higher compared to the genome assemblies of other vertebrates (Figure 5, also see Supplementary Figure 1).

TRs cause fragmentation of non-PacBio based assemblies

To investigate the possible genomic features associated with gaps in APLILM, CA454ILM, CA454PB, NEWB454, gadMor1 and gadMor2 assemblies, we mapped the contigs from each assembly to gadMor2 and categorized the intersections between the contig termini (i.e. the positions of the terminal nucleotides of each contig) and different annotated features such as

Table 4 The repeat content of of the Atlantic cod genome assembly.

Repeat	Number of elements	Coverage (Mbp)	Coverage (%)
LINEs	64,344	18.4	2.86
LTR elements	81,087	22.3	3.47
DNA elements	269,835	46.5	7.23
Unclassified	215,676	59.2	9.21
Total interspersed repeats		147.1	22.86
Tandem repeats	582,198	51.2	7.96

Groups of elements covering less than 1 % of the genome assembly are not shown.

Table 5 Overview of tandem repeat statistics.

Assembly	Total size assembly (Mbp)	Number of TRs	Mean length (standard deviation) (bp)	Density of TRs (% of assembly)
gadMor1	832	970,798	56.50 (45.17)	8.75
ALPILM	660	530,801	49.64 (53.64)	5.41
NEWB454	656	601,043	60.35 (62.72)	7.01
CA454ILM	647	921,184	73.43 (97.89)	10.2
CA454PB	682	890,967	86.01 (130.64)	10.6
gadMor2	643	876,691	84.32 (121.86)	10.9

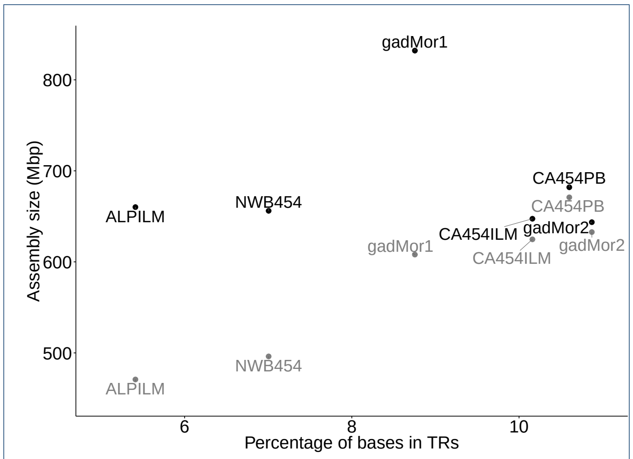

Figure 3 The density of TRs and the size of the assembly for different cod assemblies. The different assemblies (black) are all similar in size, around 650 Mbp , with the exception of the much larger gadMor1, while the amount of sequence in contigs in the different assemblies (grey) differ substantially. The vertical distance between pairs of points for each assembly equals the amount of sequence in gaps.

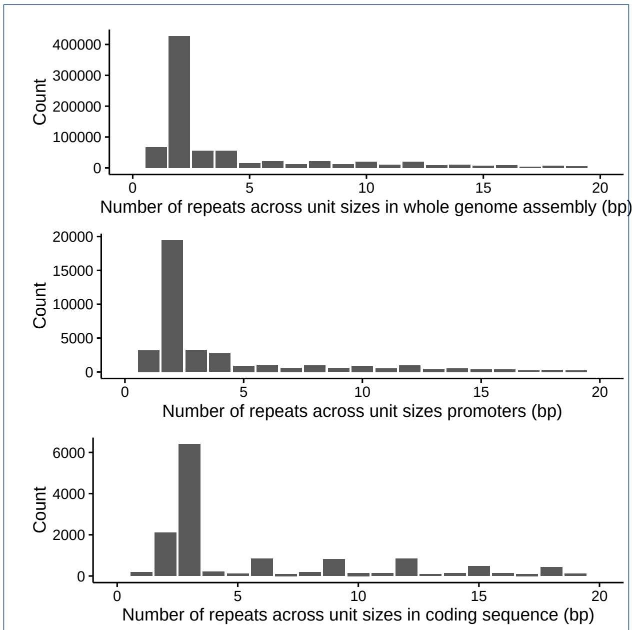

Figure 4 The number of tandem repeats categorized based on unit size. Only TRs with unit size 1-20 bp are shown. A unit size of one indicates a mononucleotide tandem repeat, two a dinucleotide, three a trinucleotide, repeats etc. The horizontal axis denotes the unit sizes of the repeat, while the vertical axis shows the count of the particular repeat.

SNPs, INDELs, TRs, transposons and lack of sequence coverage.

For gadMor2, contig termini overlap most prominently with regions lacking read coverage by any sequencing technology, and annotated transposons. The CA454PB shows the same pattern, albeit with a larger fraction of contig termini not overlapping any annotation, suggesting that these contigs end in large repeats not resolved by any assembly. For the other assemblies, the largest fraction of contig termini overlap with TRs, at percentages that are significantly higher (40+ %) than the fraction of the gadMor2 assembly annotated as such repeats (10.9 %, Table 5). All assemblies

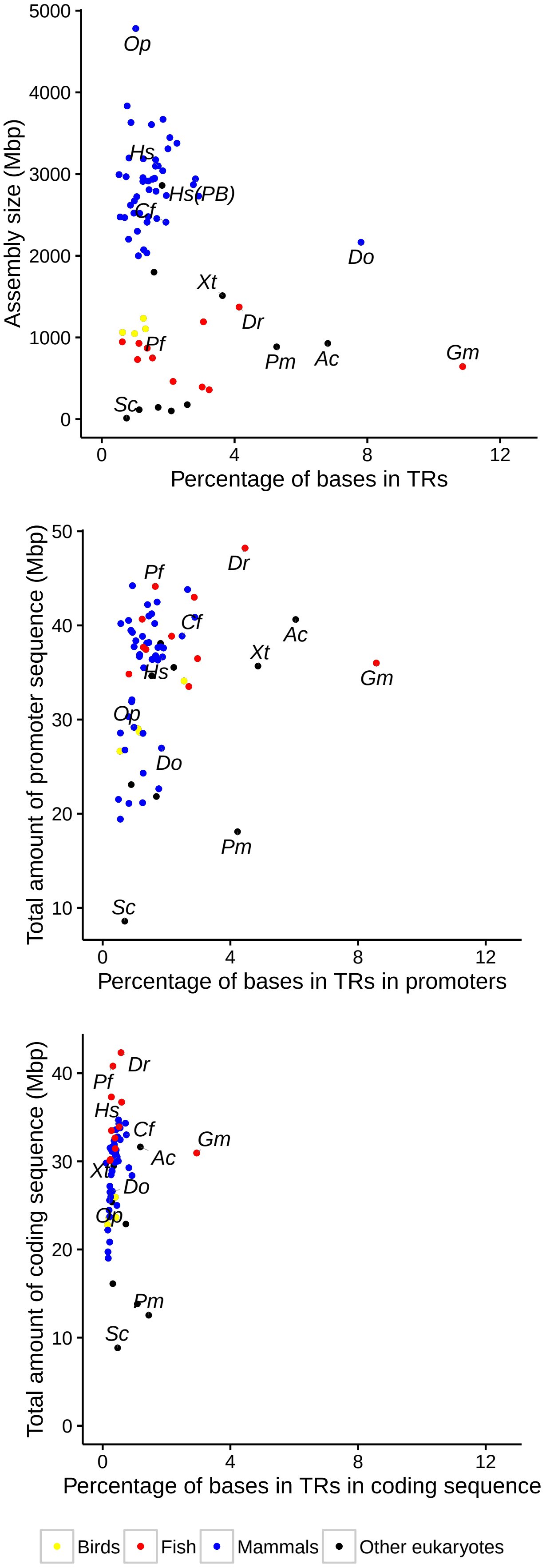

Figure 5 The density of tandem repeats in genome assemblies, promoters and coding regions. The assemblies shown here are from Ensembl release 81, excluding gadMor1, plus a human genome based on PacBio data, the California sea hare *Aplysia californica* and gadMor2 (n = 71). The panels show the density (percentage of bases) of TRs in the whole assembly, coding regions and promoter regions, respectively. The human PacBio assembly is not included in the gene and promoter analysis because it has no annotation, and the opossum is lacking for technical limitations. The species marked are *Oc* (*Ochotona princeps*, pika), *Hs* (*Homo sapiens*, human), *Hs(PB)* (*Homo sapiens*, human, PacBio based assembly), *Cf* (*Canis familiaris*, dog), *Do* (*Dipodomys ordii*, kangaroo rat), *Xt* (*Xenopus tropicalis*, frog), *Pf* (*Poecilia formosa*, Amazon molly), *Dr* (*Danio rerio*, zebrafish), *Pm* (*Petromyzon marinus*, lamprey), *Sc* (*Saccharomyces cerevisiae*, yeast), *Ac* (*Aplysia californica*, California sea hare) and *Gm* (*Gadus morhua*, Atlantic cod, gadMor2).

display difficulties in spanning transposable elements (Figure 6, Supplementary Figure 2).

Heterozygous TRs

We used lobSTR [64] to investigate the occurrence of heterozygous TRs (i.e., different repeat length between the same locus on the homologous chromosomes) in the sequenced cod's genome. lobSTR analyses TRs with unit length of 1-6 bp (i.e., STRs), and uses Tandem Repeats Finder (TRF) [65] to detect them in the genome assembly. lobSTR both annotates the STRs and discovers variation in STR length. In the sequenced individual, lobSTR annotated 980,400 STRs that passed filtering (1,182,796 in total, see Methods), of which 47,718 were heterozygous.

Compared to Phobos (which annotated 640,938 TRs of units 1-6 bp), lobSTR annotated almost twice as many STRs, and the distributions of the lengths of STRs between the two programs are quite different (Supplementary Figure 3), with lobSTR identifying relatively short STRs, and Phobos annotating relatively long STRs. Given that lobSTR is based on alignment of the 100 bp read length Illumina reads, and since the average length of a TR is 84.32 bp (Table 5), lobSTR's ability to detect heterozygous STRs is limited to around 45 bp [66]. As an alternative, we used the intersection between TRs annotated by Phobos and indels annotated by either FreeBayes (using Illumina reads, 169,635 intersections) or PBHoney (using mapped PacBio reads, 43,521 intersections). The union of these two sets comprised 145,435 indels in the 640,938 STRs (1-6 bp unit size) annotated by Phobos, about three times as many as annotated by lobSTR alone. For TRs of unit sizes 1-50 bp, there are 183,898 indels in 876,691 TRs (21 %). Altogether our results indicate that at least one-fifth of the TRs in the sequenced individual are heterozygous.

Tandem repeats in genes and promoters

We investigated the intersection of tandem repeats and coding regions, and found 17,800 coding regions in 7,372 genes intersecting with a TR. 2,094 TRs are heterozygous (based on the union of mapped PacBio and Illumina data), i.e., 12 %. These TRs are found in 1,514 genes (i.e. 6.5 % of annotated genes).

We additionally investigated the 2 kbp upstream of annotated genes (Figure 5). Of the 42,244 TRs identified in these promoter regions, 8,516 (19 %) have an indel annotated based on the union of PacBio and Illumina data.

Discussion

An improved genome assembly for Atlantic cod

We here present a new and significantly improved version of the Atlantic cod genome assembly with successful integration of different sequencing technologies. The final assembly (gadMor2) was created using a novel reconciliation method, aimed to combine the strengths of four separate assemblies to an integrated assembly that is maximized with regards to desired metrics, i.e. contig length, scaffold lengths, gene content and accordance with read data (Table 1). The individual assemblies used for the reconciliation were based on different combinations of sequencing technologies and assembly programs and varied widely in quality along the metrics studied. Importantly, the inclusion of the long PacBio reads, which spanned many more repeats than other sequencing technologies, resulted in an assembly (CA454PB) with a contig N50 an order of magnitude larger than the other assemblies, and this contributed directly to the large contig N50 of the final assembly. To our knowledge, the specific approach used in generating CA454PB, where the raw, uncorrected PacBio reads were first trimmed and then assembled without correction, together with Illumina and 454 data (see Methods), has not been described previously. A similar approach was used in generating one assembly for Atlantic salmon (see Supplement to [9]), but the sequence in that assembly did not contribute to the final assembly. End-sequenced BAC (Bacterial Artificial Chromosomes) libraries provide long-range information in the 100 kbp range, and we have such sequences available for Atlantic cod [5]. The insert size distribution of the BAC-end library was bi-modal (Supplementary Figure 10 in [5]), which is not handled properly in the Celera Assembler. We therefore included this data in the Newbler assembly (NEWB454) only, which contributed this assembly having the largest N50 scaffold of the original assemblies. The assembly using a combination of 454 and Illumina sequencing reads (CA454ILM) was the most complete assembly with regards to genes as found by the assembly validation tools CEGMA and BUSCO. While the available Illumina sequencing read datasets did not exactly match the recommendations for ALLPATHS-LG [31], the resulting assembly (ALPILM) performed better than gadMor1 with regards to N50 contig and scaffold metrics. Still, this assembly did contribute in the assembly reconciliation process, resulting in some longer scaffolds. Our results illustrate a dilemma for obtaining high-quality genome assemblies: different combinations of datasets and software using algorithms optimized for certain characteristics of the datasets yield assemblies that are of good quality on different combinations of desired quality criteria, but hardly ever on all [30]. Assembly reconciliation helps solve this issue [41], but even our integrated assembly does not rank best on every single metric evaluated. Further improvements in sequencing tech-

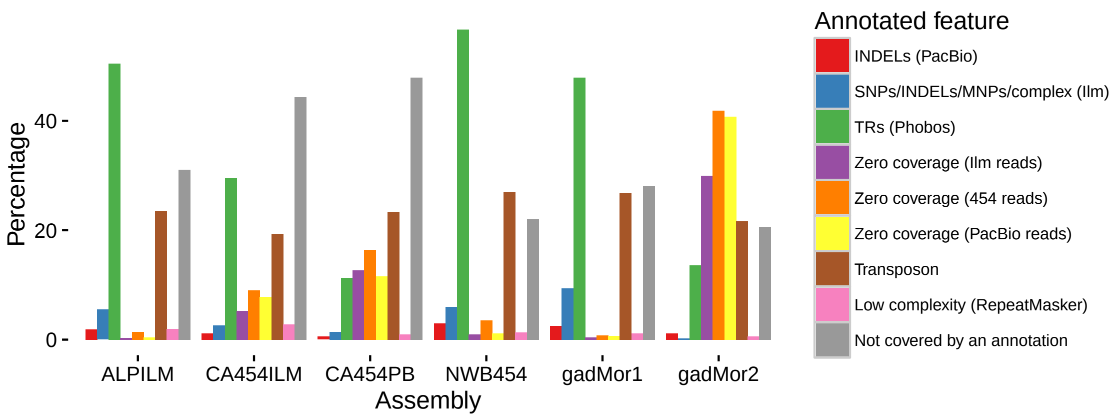

Figure 6 The intersections between contig termini and different annotated features The percentage of contig termini (the position of the terminal nucleotides of each contig) intersecting different annotations of the genome.

nology and assembly algorithms are necessary to make genome assembly a solved problem.

Due to the fragmented nature of the first version of the Atlantic cod genome, gadMor1, gene-models were reconstructed during annotation using information from the stickleback gene annotation (i.e., ordering and orienting the contigs based on stickleback gene models), and was manually curated (Supplementary Note 17 in [5]). In contrast, the gadMor2 gene models were automatically annotated directly on the genome assembly. This automated annotation did not annotate pseudogenes, in contrast to the manual curated annotation for gadMor1. The difference in annotation might explain why the CEGMA validation results are slightly lower for the new reference genome since well-annotated gene models in stickleback would be transferred to gadMor1 (Table 1). The gadMor2 assembly shows fewer indications of potential assembly errors as detected by FRC^{bam} and by comparison to the linkage map, but more according to the REAPR program. This difference is associated with longer contigs and scaffolds in gadMor2, which enabled REAPR to estimate more long-range errors. The predicted transcriptome is larger in gadMor2 (Table 2), although more genes are found with BUSCO in the gadMor1 predicted transcriptome. The genes BUSCO is designed to detect are often short (as conserved genes are often short [67]), which means they are more likely put together properly in the gene-model optimized gadMor1 assembly, since longer genes would be more likely to be fragmented.

Causes of fragmentation of cod assemblies

To understand the fragmented nature of gadMor1, we first focused on the rate of heterozygosity, as substantial differences between the homologous chromosomes of diploid organisms can fragment an assembly [53]. We compared the heterozygosity rate of the gadMor2 genome assembly (same individual as gadMor1) to three other fish with genomes for which such data is available, i.e. the miiuy croaker [51], three-spined stickleback [50] and Atlantic herring [52], and to the seq squirt *Ciona savignyi* [53], a species with extremely high heterozygosity (Table 3). The genomes for the fishes have been assembled to high contiguity (Table 3). Although a direct comparison might be confounded by the differences in population structure and the individuals chosen for sequencing (in addition to different datasets and programs used [68]), and with more uncertainty connected with calling indels correctly than with SNP calls [69], there are substantial differences between the different species. gadMor1 had a N50 contig size of 2.3 kbp 1, substantially smaller than even *Ciona savignyi* which has an order of magnitude higher SNP rate than Atlantic cod. While the species with higher SNP rates seem to have smaller N50 contig size (disregarding cod), the sequencing and assembly strategies for the different organisms vary. For gadMor1, may have had some impact on the fragmentation (Figure 6), but is not the main explanation.

Different combinations of sequencing technology and assemblers vary in their proportion of TRs present in the resulting genome assembly (Figure 3). Assemblies with higher density in TRs also have more sequence in contigs (i.e., less sequence in gaps), indicating that TRs are more completely assembled. The more fragmented assemblies (ALPILM, NEWB454 and gadMor1) have a lower density of TRs and shorter TRs on average, suggesting that TRs led to fragmentation of the assembly (Table 5). Indeed, these assemblies have a much higher proportion (40+ %) of contig termini intersecting TRs (Figure 6) than the TR density of 10.9 % in gadMor2 (Table 5). Only CA454PB and (the largely CA454PB derived) gadMor2 have about 10 % of their contig termini intersecting TRs. The remaining gaps in CA454PB and gadMor2 are associated with a lack of sequence coverage and transposons longer than the PacBio read lengths (Figure 6). This illustrates the importance of the availability of the PacBio reads, which was the only read type able to span the multitude of TRs in the genome. As illustrated in Figure 2, gadMor2 has a much higher contiguity, while a large fraction of gaps in gadMor1 are flanked with TRs. Thus, our approach taken to assemble the genome has addressed the fragmentation that affected the gadMor1 assembly. In conclusion, the high occurrence of TRs in the cod genome has caused the fragmentation of gadMor1 and all assemblies except CA454PB and (the largely CA454PB derived) gadMor2. Without the inclusion of reads obtained from the PacBio technology, or similar sequencing technologies that can span long tandem repeats, assembly of genomes with a high density of TRs, such as the Atlantic cod's, to a high sequence contiguity will be significantly more challenging.

The Atlantic cod genome reveals an extraordinary high density of TRs

We have confirmed and extended previous results showing high genomic densities of STRs in Atlantic cod [63, 70] in a sample comparing 68 genomes of eukaryotes (mostly vertebrates, Figure 5). While most of the species studied have fewer than 2.5 % of bases in TRs, California sea hare, kangaroo rat and Atlantic cod have more than 6 % bases in TRs. Atlantic cod has by far the highest density (amount of sequence in TRs) and frequency (the rate of TRs, Supplementary Figure 1) of tandem repeats in the whole genome assembly, coding regions and promoters, with only California sea hare having a higher frequency (but not density) of TRs in promoter regions, that is, more but shorter repeats.

Potential role of TRs in evolutionary processes in Atlantic cod

The mutation rates of TRs, and especially STRs, are orders of magnitude higher than that of other genomic sequences [19, 71, 72]. In the sequenced individual, we find that a fifth of the annotated TRs are heterozygous, with somewhat lower proportions in promoters (19 %) and coding regions (12 %). These results are based on the mapping of Illumina and PacBio reads, but are likely underestimations. Most of the TRs in cod have a short repeat unit, and these mutate by adding or removing, for instance, two nucleotides in the case of dinucleotide repeats. Small differences between two long alleles of a TR would likely not be captured in the analyses in this work, because the Illumina reads would not map well to these [15, 73], and the PacBio reads might not give sufficient resolution.

In humans TRs are best known for connection with multiple diseases such as Huntington's Disease [74]. In other species, variability (multiple alleles at a locus within a population) in TRs in promoter regions has been connected with diverse phenomena as behavior in voles [75] to skull form in dogs [76]. In both *Saccharomyces cerevisiae* and humans, some promoter regions contain TRs [77, 78], for which variation in length has been linked to variation in expression [78, 79]. TRs in promoter regions might also contribute to expression divergence in great apes [80] and might be connected to speciation in primates [81]. There is also variability in TRs in genes leading to functional variation as in *Saccharomyces cerevisiae*, where TRs in cell-wall genes underlie variation that causes alterations in phenotype, with different genotypes have differences in adhesion, flocculation or biofilm formation [82]. Further, in Hawaiian mints, variation in a gene coding for a flowering time protein is connected with colonization and radiation of the plant, with longer versions of the gene existing in younger populations and is suggested to contribute to morphological change and speciation [83]. It is interesting to note that Atlantic cod has a higher frequency of TRs than these species in both promoters and coding regions (Supplementary Figure 1).

The sequenced individual was from the North-East Arctic cod population, the largest cod population in the world [84], with a large effective population size [85]. Extrapolating the high mutation rate of TRs, and the observed level polymorphism in this single individual, suggests that most TRs are polymorphic at a population level. These polymorphic TRs will contribute substantially to standing levels of genomic variation in Atlantic cod populations within and in the vicinity of genes.

Conclusions

Atlantic cod has an extraordinary amount of tandem repeats compared to other species. This repeat content has led to complications in assembling the genome, which has been overcome with the usage of the long PacBio sequencing reads and reconciliation of multiple assemblies. The large amount of tandem repeats are likely to have an evolutionary impact, since they should result in a large amount of repeat-associated genetic variation in Atlantic cod populations. It remains to be investigated how cod populations evolve under variable environmental conditions with respect to TRs, and whether selection for repeat variation can lead to rapid evolutionary adaptations.

Methods

Sequencing

All read datasets originated from DNA samples from the same individual fish, designated NEAC_001, wild-caught and of the North-East Arctic population, see Supplementary Table 1.

Roche/454 reads were sequenced as described previously [5]. The Roche/454 software gsRunProcessor version 2.6 was used to redo basecalling for all sequencing runs generated for the NEAC_001 sample [5].

180 bp insert size and 300 bp insert size libraries were constructed with Illumina DNA paired end sample preparation reagents and sequenced at the Norwegian Sequencing Centre. The 5 kbp insert size libraries were prepared with the Illumina Mate Pair gDNA reagents and sequenced at the McGill University and Génome Québec Innovation Centre. All Illumina libraries were sequenced on the HiSeq 2000 using V3 chemistry 100 bp paired end reagents.

PacBio SMRT sequencing was performed on PacBio RS instrument (Pacific Biosciences of California Inc., Menlo Park, CA, USA) at Norwegian Sequencing Centre (www.sequencing.uio.no/) and at Menlo Park. Long insert SMRTbell template libraries were prepared at NSC (10kb insert size) and Menlo Park (22 kb insert size) according to PacBio protocols. In total, 147 SMRT cells were sequenced using C2 and XL polymerase binding and C2 and XL sequencing kits with 120 min acquisition. Approximately 7.6 Gb of library bases were produced from 10 kb SMRTbell libraries sequenced on 102 SMRT cells using C2/C2 chemistry (average polymerase read length of 3 kb). The 22 kb SMRTbell library was sequenced using C2/XL (22 SMRT cells, average polymerase read length of 4.5 kb) and XL/XL (23 SMRT cells, average polymerase read length of 5 kb) chemistry producing 5.5 Gb of library bases.

Assembly

An overview of the usage of different sequencing data in the different assemblies is in Supplementary Table 1.

ALLPATHS-LG assembly, ALPILM

An ALLPATHS-LG [31] assembly was created using only the Illumina reads. Paired end 100 bp Illumina reads from a 180 bp insert size library were input as fragment reads, while paired end 100 bp reads from a 300 bp insert library and 100 bp reads from a 5k mate pair library were input as jumping reads. Only half of the fragment reads were used in the assembly (Supplementary Table 1), selected as an option, to have the recommended coverage. The release R48639 of ALLPATHS-LG was used.

Newbler assembly, NEWB454

Newbler version 3.0 was used to assemble the 454 sequencing data together with BAC-ends previously generated for [5], with the options ”-large -het -repfill -sio -info -a 0”. In contrast to the Newbler assembly done for the first version of the Atlantic cod genome [5], we did not filter out 454 reads consisting entirely of short tandem repeats, as newer versions of the Newbler program are better able to deal with these reads.

In its output, Newbler gives a file with all scaffolds, including all unscaffolded contigs longer than 2 kbp, and a separate file with all contigs, regardless of their inclusion in a scaffold. Using BLAT version 3.5 [86] we mapped the flanking sequences of SNPs in the linkage map (personal communication, Sigbjørn Lien) (n=9355) to all contigs. For each mapped SNP, the longest contig to which it mapped was added to the primary output, with the rationale that sequences with SNPs should be included in the assembly. The final assembly thus contain all scaffolds, all contigs longer than 2 kbp and the longest unplaced contigs with a mapped SNP.

Celera Assembler assembly based on 454 and Illumina reads, CA454ILM

Celera Assembler's meryl (SVN snapshot dated 2nd of April 2013) [32] was used to count kmers in the two paired end Illumina read libraries, of 180 bp and 300 bp insert sizes and of length 100 bp.

FLASH version 1.2.3 [87] was used to merge the overlapping reads from the 180 bp library using default options.

The merTrim program, also from Celera Assembler, was used to correct Illumina reads by changing infrequent kmers to frequent kmers: starting from the first (last) frequent kmer in a read, if the next (previous) kmer is infrequent, then the most recently added base must be an error. To correct it, the three substitution changes are tested; if all kmers spanning this base are now frequent, the change is accepted. If not, the four insertion and one deletion changes are tested; likewise, if all kmers spanning this change are now frequent, the change is accepted. Otherwise, the base is left unchanged. Finally, the read is trimmed to the largest region with all kmers designated as frequent kmers.

Celera Assembler was used to remove duplicate reads from the 300 bp and 5 kbp Illumina reads libraries with its run runCA-dedupe pipeline.

All 454 reads were converted from .sff files to .fastq and .frg files using Celera Assembler's sffToCA with options "-linker flx -linker titanium -insertsize ins_size std_ins_size -trim chop -libraryname lib_name -output output_name", with insert sizes and standard deviations at 1100, 320; 1230, 350; 1440, 440; 1760, 470; 2650, 700; 7000, 1900; 19000, 4750 for the different sequencing libraries increasing in insert size (Supplementary Table 1). The insert sizes and standard deviations were gathered from the output of a Newbler assembly which calculated this.

The 454 reads were error corrected using the merTim program, as above, and trimmed as described in Prüfer et al. [88], which removes duplicated pairs of reads, error-prone ends of reads, reads with sequence not confirmed by other reads and chimeric reads. Because the insert length distribution of the paired reads from the 20 kbp 454 mate pair library showed a bimodal distribution (Supplementary Figure 4, in [5]), and because Illumina mate pair libraries contain contamination with pair of reads with the opposite orientation, the scaffolds from this assembly were used to filter out reads from the 20 kbp 454 library and the 5k Illumina library by mapping the reads to the scaffolds using BWA-MEM [54], and removing any pair of reads that mapped closer than 10 kbp and 2 kbp, respectively.

After the error correction steps, all 5 kbp mate pair Illumina reads, 6x coverage of the 300 bp insert size Illumina reads and 25x of the merged 180 bp insert size Illumina reads were assembled together with all the 454 reads. Seqtk [89] from November 2012 was used to extract these reads.

The assembly used this spec file (only non-default options shown):

```
unitigger = bogart
batThreads = 64
doExtendClearRanges=0
doToggle = 0
cgwMergeFilterLevel = 2
cgwMinMergeWeight = 2
```

Contigs from Celera Assembler's degenerate contig file, normally excluded from scaffolds, were added to the assembly if they contained flanking sequence from a SNP used on in the SNP-chip as described above for the Newbler assembly.

Celera Assembler assembly based on PacBio, 454 and Illumina reads, CA454PB

All processing of Illumina and 454 reads were redone as described above, using Celera Assembler 8.1.

Filtered subreads of PacBio reads were trimmed using Celera Assembler 8.2 alpha with this spec file (only non-default options shown):

```
stopAfter = overlapBasedTrimming
merSize = 16
merThreshold = 0
merDistinct = 0.9995
merTotal = 0.995
ovlErrorRate = 0.40
ovlMinLen = 500
doFragmentCorrection = 0
```

Assembly below was run with this spec file (only non-default options shown):

```
merSize = 16
merThreshold = 0
merDistinct = 0.9995
merTotal = 0.995
doOBT = 0
doDeDuplication = 0
ovlErrorRate = 0.40
frgMinLen = 100
ovlMinLen = 100 #Changed for each overlaps
    between each technology, see below
doFragmentCorrection = 0
unitigger = bogart
utgGraphErrorRate = 0.300
utgGraphErrorLimit = 32.5
utgMergeErrorRate = 0.35
utgMergeErrorLimit = 4
utgBubblePopping = 1
utgErrorRate = 0.40
utgErrorLimit = 25
batThreads = 16
cgwDemoteRBP = 0
cgwErrorRate = 0.40
doExtendClearRanges = 0
doToggle = 0
cgwMergeFilterLevel = 2
cgwMinMergeWeight = 4
cnsErrorRate = 0.40
doUnitigSplitting = 0
cnsMaxCoverage = 40
cnsReuseUnitigs = 1
```

The assembly contain all paired 454 reads, 25x of merged reads from the 180 bp insert size Illumina library and the trimmed PacBio reads, and was run with Celera Assembler 8.2 alpha. To accommodate vastly different error rates between the Illumina/454 and PacBio reads, overlaps were computed using a different percentage maximum allowed error (inverse of percentage identity) cutoff for each pair of technologies being overlapped. Overlaps between Illumina and 454 reads were computed to a maximum of 6 % error and minimum overlap of 100 bp; overlaps between an Illumina/454 read and a PacBio read were computed to a maximum of 20 % error, also with a minimum overlap of 100 bp; overlaps between two PacBio reads were computed to a maximum of 40 % error and minimum overlap of 1000 bp. The bogart unitig construction algorithm will pick, for each read end, the longest overlap and use only those for constructing initial unitigs, similar to the BOG algorithm in [32]. Bogart uses clusters of partially aligned reads (discovered via pre-computed overlaps) to detect junctions between repeat and non-repeat sequence. If a detected repeat is spanned by either a read or a mate-pair, the repeat is left intact, otherwise, the unitig is split into at least three pieces: one for each side of the repeat, and at least one for the repeat itself.

The rest of the assembly process was run as normal, aside from much higher error rate acceptance at all steps and a non-default selection of unique unitigs. Because PacBio reads confuse Celera Assembler's classification of unique unitigs (which can be used as seeds for creating contigs) and non-unique unitigs (often repeats that could be placed several times in the assembly), we ran the classification tool markRepeatUnique by hand, specifying that unique unitigs could not have a single reads spanning more than 90 % of its length, up to 15 % of the unitig could have a depth of only 3 reads, and must have had at least 200 reads and be at least 10,000 bp long. Command:

Degenerate sequences that either contained a SNP (as described earlier) or a gene found with CEGMA version 2.4.010312 [38, 67], were added to the assembly output.

Pilon and PBJelly

All four assemblies described above were processed with PBJelly (SVN snapshot 23rd September 2014) [34], a tool that maps PacBio reads back to the assembly and uses them to close gaps both between and within scaffolds. Protocol.xml:

```
<jellyProtocol>
 <reference>genome.fasta</reference>
    <outputDir>output</outputDir>
    <blasr>-minMatch 12 -affineAlign -
        minPctIdentity 75 -bestn 1 -nCandidates
        10 -maxScore -500 -nproc 16 -
        noSplitSubreads</blasr>
    <input>
              <job>pacbio_reads.fastq</job>
    </input>
</jellyProtocol>
```

Commands used:

```
Jelly.py setup Protocol.xml -x "--minGap 20"
Jelly.py mapping Protocol.xml
Jelly.py support Protocol.xml
Jelly.py extraction Protocol.xml
Jelly.py assembly Protocol.xml
Jelly.py output Protocol.xml -x "-m 3"
```

Pilon version 1.9, a program to automatically improve assemblies [33], was applied to both the original and the PBJelly version of the assemblies, using all 454 reads, the reads from the 300 bp and 5 kbp insert size Illumina libraries, mapped with BWA-MEM 0.7.9a and sorted by samtools 0.1.19:

```
bwa mem genome.fasta -M reads.fastq 2> log.err |
    samtools view -buS - | samtools sort -
    reads_mapped.sorted
```

Pilon options were (not showing all the libraries):

```
java -Xmx500G -jar pilon-1.9.jar --genome genome.
    fasta --frags paired_reads.sort.bam --jumps
    paired_reads.sort.bam --unpaired
    unpaired_reads.sort.bam --changes --diploid
    --output genome_pilon
```

And the reads from all PacBio libraries, mapped with blasr from SMRTanalysis 2.2.0 and sorted by samtools 0.1.19:

```
sawriter genome.sa genome.fasta
blasr -sa genome.sa reads.fastq genome.fasta -
    bestn 2 -sam -clipping soft -minMatch 12 -
    affineAlign -nCandidates 8 -minPctIdentity 75
     -out reads.sam -nproc 16
cat reads.sam | samtools view -buS - | samtools
    sort - reads.sort
```

This resulted in four different versions of each assembly: the original; one processed with PBJelly; one processed with Pilon; one and processed with both PBJelly and Pilon. Based on the results of the validation tools against applied to all versions of the assemblies (see below), one version of each assembly was chosen for merging, the versions of ALPILM, NEWB454 and CA454PB after application of both PBJelly and Pilon and the version of CA454ILM after application of Pilon only.

Validation

To evaluate assembly quality, several validation tools were applied. Both REAPR [36] and FRC^{bam} [35] use paired Illumina reads to evaluate an assembly, giving a measure of the number of potential errors. Instead of using the raw reads, we used error corrected reads dumped from the ALLPATHS-LG assembly, reducing the running time of both the alignment step and the tools themselves.

Isoblat was used to determine how much of the Newbler transcriptome of 454 and Sanger reads was aligned to the different assemblies [37]. It was run with default options.

CEGMA is a tool that annotates 458 highly conserved genes in an assembly, and it can be used to assess the completeness of the genome assembly [38, 67]. Version 2.4 was applied to all different versions of the assemblies.

BUSCO is similar to CEGMA in that it assesses the completeness of a genome by trying to find a set of universal single-copy orthologs [39]. In this study, we used the actinopterygii specific set of 3698 genes to investigate the completeness of the assemblies generated here.

9355 of SNPs have been used to produce a linkage map (personal communication, Sigbjørn Lien). We used blat_parse.py to compare the linkage map to different assemblies to evaluate the completeness and long-range correctness. Briefly, this involved mapping the flanking sequences of the SNPs to the assembly using BLAT version 3.5 [86] and options ”-noHead -maxIntron=100 genome.fasta flanking_sequences.fasta” and then parsing the output file while comparing with the order of the SNPs in the linkage map. A conflict with the linkage map is defined when a sequence had SNPs mapped to it belonging to more than one linkage group. Some SNPs mapped equally well to more than one linkage group, and these were excluded since we could not confidently judge which mapping was correct.

Merging of assemblies

Each assembly was aligned against itself using nucmer [90], and any sequences fully contained in another sequence with more than 98 % identity were removed. Scaffolds were split with a split_asm_lg.py (available on the github repository together with the other scripts mentioned in this section) if they conflicted with the linkage map. A scaffold in conflict is split into three pieces, from the start for the scaffold following one linkage group to the last basepair in the flanking sequence of the last SNP in that linkage group, and from the first basepair in the flanking sequence of the first SNP in another linkage group. The middle piece is not used since we do not know where exactly the transition from linkage group to another happens. Sequences shorter than 1000 bp were removed to better facilitate the whole assembly alignment process.

The four assemblies selected for merging were aligned together using Mugsy. Mugsy uses nucmer from the Mummer package [90] to find similar sequence in different assemblies and subsequently refines the alignment. It outputs a MAF (Multiple Alignment Format) file, consisting of blocks of multiple alignments with information where exactly in the sequences the alignment is (starting at 100 bp and ending at 300 bp in scaffold X in assembly Y for instance), which was parsed by merge_asms.py. Based on validation criteria described above, one assembly was chosen as the skeleton (CA454ILM), and a second assembly was chosen as the sequence contributing part (CA454PB). The CA454ILM assembly was chosen as skeleton because it was the most complete with regards to genes, and CA454PB was chosen as sequencing contribution assembly was chosen because it had the least gaps. A first pass through the alignment blocks of the first assembly was used to close gaps using the sequences from the CA454PB assembly, or the sequence in each alignment block with the least amount of missing bases. A second pass through the alignment blocks of the first assembly tried to connect scaffolds from the first assembly (CA454ILM) using scaffolds from other assemblies spanning two scaffolds in CA454ILM. Mugsy was run with these options:

```
mugsy --directory output_folder \
CA454ILM_pilon_dedup_98_split_min_1000.fasta \
NEWB454_pbjelly_pilon_dedup_98_split_min_1000.
    fasta \
ALPILM_dedup_98_split_min_1000.fasta \
CA454PB_pbjelly_pilon_dedup_98_split_min_1000.
    fasta \
-nucmeropts "-l 150 -c 1000 -g 90000" -c 500 -
    fullsearch > mugsy.out 2> mugsy.err
```

We mapped all paired Illumina and 454 reads to the assembly with BWA-MEM 0.7.9a, and used the scaffold module from SGA [43] to scaffold the merged assembly, increasing N50 scaffold from 850 kbp to 1.15 Mbp. Pilon was then applied using all reads excluding PacBio and the 180 bp insert size Illumina library.

Anchoring to linkage map

Finally, the scaffolds were ordered into linkage groups based on linkage data (personal communication, Sigbjørn Lien) with 100 Ns between two adjacent scaffolds using order_orient_scaffolds.py. Scaffolds with only one SNP kept their existing orientation, while scaffolds with more than one SNP were reverse complemented if more than half the SNPs suggested this. The numbering of the linkage groups is according to Hubert et al. [91].

Transcriptome assemblies

We obtained transcriptome datasets from three different sequencing technologies, Illumina, 454 and PacBio, from a variety of tissues and different stages. Three different transcriptome assemblies were created: (i) based on assembly of the Illumina reads using Trinity [92]; (ii) assembly of the 454 reads using Newbler [93]; and (iii) clustering the long full-isoform PacBio reads using SMRT-Analysis [94].

Trinity with Illumina reads

RNA-seq sequencing data used in Penglase et al. [95] (from larvae at different stages and feeding regimes, the Sequence Read Archive (SRA) at NCBI with accession ID: SRP056073) were obtained and adapters and all bases with less than 20 in Phred quality score were removed with cutadapt 1.5 [96]. Trinity version r20140717 [92, 97] was run with the normalize_reads option turned on. 654948 transcripts were assembled. Abundance estimates commands:

```
align_and_estimate_abundance.pl \
--transcripts trinity_out_dir/Trinity.fasta \
--seqType fq \
--est_method RSEM \
--aln_method bowtie --trinity_mode --
    prep_reference \
--left read1.fq --right read2.fq --thread_count
    16
```

The script filter_fasta_by_rsem_values.pl distributed with Trinity was used to filter the transcript assembly based on abundance, where only transcripts with fragments per kilobase of transcript per million mapped reads (FPKM) of at least 0.05, and a transcript abundance of at least 1 % of the parent gene's abundance were kept, resulting in 59,379 transcripts.

Newbler with 454 and Sanger reads

The transcriptome 454 and Sanger reads used in Star et al. [5] (the different tissues listed in Supplementary Table 2 in [5]) were combined with Sanger reads from Kleppe et al. [98], and assembled with Newbler 3.0 using the options -cdna and -vt with these primers:

```
>5prime
CTACTAGACCTTGGCTGTCACTCA
>3prime
TCGCAGTGAGTGACAGGCTAGTAG
>1
TACAGGCCATTACGGCCGGGG
>2
TTTTTTTTTT
>3
TTTTTTTTTTTTTTTTTT
```

The assembly resulted in 79,025 transcripts.

IsoSeq on PacBio reads

Equal amounts of RNA were isolated from pool of unfertilized eggs and 20, 30, 45, 60 and 90 days post hatch. This was pooled and three size-selected fractions based on agarose gel-electrophoresis of RNA were created and sequenced on the Pacific Biosciences RS: 1-2 kbp, 2-3 kbp and 3-6 kbp using P6v2-C4 chemistry [94]. Using SMRT Portal, reads-of-insert were first created for each fraction, and isoform prediction and polishing by Quiver were performed according to the manufacturer's instructions. For the fraction 1-2 kbp, 10,738 high quality isoforms were predicted ($\leq$ 99 % accurate sequence according to Quiver) and 2,952 low quality (<99 % accurate sequence), for the 2-3 kbp fraction 15,688 high quality and 6,898 low quality and for the 3-6 kbp fraction 13,400 high quality and 12,716 low quality transcripts. These 62,392 transcripts were merged into one fasta file and used in further analyses.

Annotation

Repeat libraries

A repeat library for MAKER gene annotation (see below) was created by running RepeatModeler [58] version 1.0.8 on the finished genome assembly with default options.

We also created a repeat library specifically for annotation of transposable elements (https://github.com/uio-cels/Repeats). First, RepeatModeler [58] version 1.0.8 was run on only the scaffolds longer than N50. LTRharvest [59] and LTRdigest [60], both parts of genometools (version 1.5.7), were used to detect LTR retrotransposons and TRIMs. LTRharvest found LTR retrotransposons with LTRs larger than 100 nt, smaller than 6000 nt and with 1500 to 25000 nt between, with a target site duplication (TSD) length of 5 nt. TRIMs were detected by lowering the LTR length requirements to a minimum of 70 nt and a maximum of 500 nt with maximum 1500 nt of internal sequence. Harvested putative LTR retrotransposons were filtered using LTRdigest, which checked for tRNA binding sites. In addition, LTRdigest used Hidden Markov Model (HMM) profiles to identify retrotransposon enzymes (from the GyDB HMM profile collection of retrotransposon specific enzymes [99]). Elements without both tRNA binding sites and a retrotransposon specific enzyme were discarded.

We used scripts provided by Ning Jiang, Megan Bowman and Kevin Childs (Michigan State University) to perform the next analyses. Only elements containing primer binding sites (PBS) and/or a polypurine tract (PPT) was kept, and only if at least half of the PBS or PPT sequence were located in the internal regions of the putative element and only if the distance between the LTRs and the PPT/PBS sequence were less than 20 bps. Elements that passed this test were subjugated to further filtering where sequence gaps of $\geq$50 nt were discarded. MUSCLE version 3.8.31 [100] was used to align flanking sequences, and elements with $\geq$ 60 % similarity in flanking sequences was excluded.

Nested LTR retrotransposons were detected by using RepeatMasker with the left LTR sequences of the putative elements and a library of transposases (from a curated library included in the software TEseeker v1.04 [101]). Consensus sequences were produced after all vs. all comparisons using BLASTN. Finally, no elements of different families shared 80 % sequence over 90 % of their length.

RepeatClassifier, which is a program included in RepeatModeler, was used to classify the elements. As LTR retrotransposons and TRIMs contain tandem repeats in their long terminal repeats, RepeatClassifier classified some elements as being tandem repeats. These elements were renamed to being LTR retrotransposons or TRIMs, while those that were classified into specific LTR families kept their new classification. TransposonPSI [97] was also run. TransposonPSI uses PSI-BLAST to detect distant homology between genomic sequences and a TE library bundled with the program. Contrary to the other programs, TransposonPSI does not output the consensus sequences of elements detected, which made it necessary to perform an additional clustering step. The output sequences were clustered using CD-HIT-EST 4.6.4 [102] with a similarity cutoff of 80 %. The relative high amount of dinucleotide repeats in the Atlantic cod genome assembly, led to a large fraction of sequence being labeled as transposons of the CACTA superfamily, as the CACTA representative in the TransposonPSI library contained a tandem repeat that spurred false alignments. Thus, elements were only named CACTA if two sources agreed in the classification, the other source being the results of a BLASTX search against the repeat peptide database provided with RepeatMasker (version 4.0.6).

As the detection tools might detect repetitive non-TE genes such as gene families, the sequences were checked for alignments (using BLASTX) with sequences in the curated protein database of UniProtKB/SwissProt [103], which was downloaded November 20th 2015. Sequences were also checked against the repeat peptide database that comes with distributions of the RepeatMasker software. Sequences with matches in the UniProtKB/SwissProt database, but not in the repeat peptide database were discarded. The BLASTX search against repeat peptides in the database also served to classify some of the unclassified elements.

Some sequences remained unclassified, and a collection of HMM profiles was downloaded from the Dfam database (Dfam.org) and HMMER3 was run using the 'nhmmer' module. This further classified some elements into LTR retrotransposons, LINEs, SINEs or DNA transposons. The *de novo* library was merged with known eukaryotic repeat sequences from RepBase [62] (version 20150807) and served as input for RepeatMasker.

Annotation with MAKER

MAKER is an annotation pipeline designed to combine the consolidated output from different *ab initio* gene finders and physical evidence (e.g. protein and RNA-seq alignments) into a set of quality scored gene models (AED score) [47, 48, 104].

A two-pass iteration with MAKER version 2.31.8 [47, 48] was performed on the final genome assembly as described at [105] and in Campbell et al. [106]. First, two *ab initio* gene finders were trained, SNAP version 20131129 [107] on the genes found by CEGMA version 2.4.010312, and GeneMark-ES version 2.3e [108] on the genome assembly itself. SwissProt/UniProtKB [103] was downloaded 9th of May 2015 (release 2015_04). MAKER was configured to use the two trained *ab initio* gene finders, the SwissProt/UniProtKB protein database [103], the RepeatModeler repeat library and three different transcriptomes, one based on 454 and Sanger data, one based on Illumina and one based on PacBio. Additional options were these:

```
genome=gadMor2.fasta
est=/path/to/newbler_transcriptome.fasta,/path/to
    /trinity_transcriptome.fasta,/path/to/
    pacbio_transcriptome.fasta
protein=/path/to/uniprot_sprot.fasta
rmlib=/path/to/repeatmodeler.fasta
repeat_protein=/path/to/te_proteins.fasta #
    provided with MAKER
snaphmm=/path/to/genome.cegmasnap.hmm
gmhmm=/path/to/GeneMark.mod
est2genome=1
protein2genome=1
keep_preds=1
single_exon=1
split_hit=20000
alt_splice=1
```

The GFF output from the first pass with MAKER was used to retrain SNAP, and to train AUGUSTUS version 3.0.2 [109, 110] with the PacBio transcriptome. A second pass with MAKER was run with the retrained SNAP, the trained AUGUSTUS and the similar set of input as above, and with these other options:

```
genome=gadMor2.fasta
est=/path/to/newbler_transcriptome.fasta,/path/to
    /trinity_transcriptome.fasta,/path/to/
    pacbio_transcriptome.fasta
protein=/path/to/uniprot_sprot.fasta
rmlib=/path/to/repeatmodeler.fasta
repeat_protein=/path/to/te_proteins.fasta #
    provided with MAKER
snaphmm=/path/to/maker1.snap.hmm
gmhmm=/path/to/GeneMark.mod
augustus_species=gadMor2
```

```
est2genome=0
protein2genome=0
keep_preds=1
single_exon=1
split_hit=20000
alt_splice=0
```

InterProScan version 5.4-47 [111] was run on the protein output of Maker, giving gene ontologies and classifying protein domains and families. The protein output was BLASTed against SwissProt/UniProtKB release 2015_12, giving putative gene names, with these options:

```
blastp -query maker.all.maker.proteins.fasta \
    -db uniprot_sprot.fasta \
    -num_threads 10 -evalue 1e-5 -outfmt 6 -
        num_alignments 1 -seg yes -soft_masking
        true \
    -lcase_masking -max_hsps_per_subject 1 \
    -out maker.uniprot-sport.blastp.1e-5.max50
```

Investigating heterozygosity

To investigate the heterozygosity of this individual of Atlantic cod, we mapped the 300 bp insert size Illumina sequencing library to the genome assembly using bwa mem version 0.7.9a with the -M option [54]. Samtools version 1.1 was used to sort the bam files.

```
bwa mem genome.fasta -M reads.fastq 2> log.err |
    samtools view -buS - | samtools sort -
    reads_mapped.sorted
```

SNP and indel calling was done on the merged bam file using FreeBayes version v0.9.14-17-g7696787 [55], and SNP and indel calls with a quality more than 20 were kept with ’vcffilter -f ”QUAL ¿ 20”’. Vcfstats was run on the resulting VCF file, giving the number of SNPs, MNPs, indels and complex regions.

We also mapped all PacBio reads using blasr from SMRT-Analysis 2.3.0, and called indels using PBHoney [57] giving all indels larger than 20 bp. This numbered 70,278.

Genome wide short tandem repeat analysis

Tandem repeats of unit size 1-50 bp were detected with Phobos version 3.3.12 [18], options set were ”-s 12 –outputFormat 0 -U 50”, i.e. requiring a minimum score of 12 for each tandem repeat, that is, the TR needed a score above 12, i.e. at least 13 mononucleotides, 7 dinucleotide, 5 trinucleotide repeat units, etc., Phobos native format as output and up to a motif, or unit, size of 50 bp. A minimum score of 12 would mean that mono-, di-, and trinucleotides repeats had to have a minimum length of 13, 14 and 15 bp to exceed the minimum score, 13, 7 and 5 repeated units respectively. A range of 1-50 bp was chosen in accordance with Mayer et al. [18]. A config file was then provided for the satstat version 1.3.12 program, giving a diverse output of file with different statistics and a gff file:

```
input  example_data.phobos
output example_stats.txt
foreach all
compute #sat %sat #units %units minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
    minmaxperfection statperfection #taxawithsat
foreach perfection
compute #sat %sat #units %units minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
    minmaxperfection statperfection #taxawithsat
foreach unitlength
compute #sat %sat #units %units minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
    minmaxperfection statperfection #taxawithsat
output example_units.txt
foreach unit
compute minmaxunitlength #sat %sat minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
output example_units10.txt
# Show only those table rows with at least 10
    tandem repeats
foreach unit 10
compute minmaxunitlength #sat %sat minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
output example_units20.txt
foreach unit 20
compute minmaxunitlength #sat %sat minmaxlength
    statlength minmaxrepeats bp/mbp corr-bp/mbp
output example_taxa.txt
foreach taxon
compute #sat %sat #units %units minmaxlength
    statlength minmaxrepeats bp/mbp rbp/mbp
compute taxoncontent
compute units&freq
output example_dist.txt
foreach all
compute #sat minmaxdistances statdistances corr-
    bp/mbp
foreach unitlength
compute #sat minmaxdistances statdistances corr-
    bp/mbp
foreach unit
compute #sat minmaxdistances statdistances corr-
    bp/mbp
output example_satdistances.txt
foreach all
compute satdistances
foreach unitlength
compute satdistances
foreach unit
compute satdistances
```

```
output example-out.gff
print-gff
exit
```

In addition, STRs were detected with lobSTR 4.0. First, TRF version 4.07b was run on the genome assembly with these options ”gadMor2.fasta 2 7 7 80 10 24 6 -f -d -h”, and the resulting gadMor2.fasta.2.7.7.80.10.24.6.dat file was converted to bed format with convert_trf_bed_lobstr.py. A lobSTR index was created with the bed file and the genome, and allelotype classified the STRs using the Illumina 300PE library previously mapped with BWA, using these options:

```
allelotype --command classify --bam
    gadMor2_300bp_raw_rg.sort.bam \
 --strinfo gadMor2_strinfo.tab --noise_model /
     path_to_lobstr/share/lobSTR/models/
     illumina_v2.0.3 \
--index-prefix gadMor2_index/lobSTR_ --out
    gadMor2 \
--filter-mapq0 --realign --max-repeats-in-ends 3
    --min-read-end-match 10
```

In addition to the different cod assemblies analyzed, we downloaded all assemblies from Ensembl release 81 (n = 68) (including Atlantic cod) and the California sea hare.

Star et al. [5] released three different assemblies, one based on Newbler, one on Celera Assembler and a gene-model optimized, annotated version of the Newbler assembly, which is the one available from Ensembl and indicated herein as gadMor1. In gadMor1 contigs were reshuffled according to stickleback proteins during annotation, which resulted in significant improvements with regards to gene model construction compared with the original assembly. In all comparisons between different cod assemblies performed for this work, we compared to the gadMor1 assembly, since it is annotated and likely the one most used.

Contig terminus analysis

Contigs from the assemblies of ALPILM, NEWB454, CA454PB, CA454ILM, gadMor1 and gadMor2 were created with the cutN -n 1 from seqtk version 1.0-r75, which cut at each gap (of at least one basepair, i.e. one or more Ns). The contigs were mapped against the gadMor2 assembly with BWA 0.7.12 and get_positions_non_soft_hard_clip.py was used to create a BED file with only the edges of contigs that map uniquely with a mapping quality of 3 or more.

The intersect option from bedtools version 2.24.0 [112] was used to find overlaps between the contig termini and indels based on PBHoney tails output, SNPs, indels, MNPs and complex regions from mapping Illumina reads (300 bp insert size) to the genome, STRs called by Phobos, lack of coverage by Illumina, 454 and PacBio reads (zero depth as determined by mapped reads and bedtools genomecov), transposons and low complexity regions from RepeatMasker.

Heterozygous tandem repeats

We used bedtools [112] 2.24.0 to find the intersecting between the indels called by FreeBayes and PBHoney, and the TRs as annotated by Phobos. Indels were filtered based on depth (at least 5 reads) and genotype (0/1, heterozygous).

```
cat gadMor2_300bp_rmdup_freebayes_single_q20.vcf
    |vcffilter -f "TYPE = del | TYPE = ins" |
    vcffilter -f "DP > 5" | grep "0/1" >
    gadMor2_300bp_indels_gt_dp5.vcf
bedtools merge -i phobos_trs.gff >
    phobos_whole_genome_trs.bed
bedtools merge -i gadMor2_honey.bed >
    pacbio_indels.bed
bedtools merge -i gadMor2_300bp_indels_gt_dp5.vcf
     > ilm_indels_gt_dp5.bed
bedtools intersect -a phobos_whole_genome_trs.bed
     -b pacbio_indels.bed >
    whole_genome_with_trs_pacbio_indels.bed
bedtools intersect -a phobos_whole_genome_trs.bed
     -b ilm_indels_gt_dp5.bed >
    whole_genome_with_trs_ilm_indels_gt_dp5.bed
cat whole_genome_with_trs_pacbio_indels.bed
    whole_genome_with_trs_ilm_indels_gt_dp5.bed |
    sort -k1,1 -k2,2n |bedtools merge >
    whole_genome_with_pbilm_indels_gt_dp5.bed
```

In the annotation of Atlantic cod, some genes were annotated that consist predominantly of TRs. Since these were in the annotation, they have some evidence in the form of protein or transcriptome alignment, and have an open reading frame. However, they seem to have no significant similarity with proteins from SwissProt/UniProtKB, and were removed based on this. This left 19,035 genes for this particular analysis.

```
cat gadMor2_maker.putative_function.domain_added.
    aed_0.5.gff |awk '{if ($3 == "gene") print $0
    }' |sort -k1,1 -k4,4n > genes.gff
grep -v unknown genes.gff > known_genes.gff
bedtools flank -i known_genes.gff -g gadMor2.
    fasta.fai -l 2000 -r 0 -s |sort -k1,1 -k4,4n
    > genes.2kb.promotors.gff
bedtools intersect -a genes.2kb.promotors.gff -b
    phobos_trs.gff |sort -k1,1 -k4,4n >
    intersect_2kb_promotor_trs.gff
bedtools merge -i intersect_2kb_promotor_trs.gff
    > intersect_2kb_promotor_trs.bed
```

```
bedtools intersect -a intersect_2kb_promotor_trs.
    bed -b pacbio_indels.bed >
    promoters_with_trs_pacbio_indels.bed
bedtools intersect -a intersect_2kb_promotor_trs.
    bed -b ilm_indels_gt_dp5.bed >
    promoters_with_trs_ilm_indels_gt_dp5.bed
cat promoters_with_trs_pacbio_indels.bed
    promoters_with_trs_ilm_indels_gt_dp5.bed |
    sort -k1,1 -k2,2n |bedtools merge >
    promoters_with_pbilm_indels_gt_dp5.bed
```

```
cat /gadMor2_maker.putative_function.domain_added
    .aed_0.5.gff |awk '{if ($3 == "CDS") print $0
    }' |sort -k1,1 -k4,4n > cds.gff
bedtools intersect -a cds.gff -b known_genes.gff
    |sort -k1,1 -k4,4n > cds_known_genes.gff
bedtools intersect -a cds_known_genes.gff -b
    phobos_whole_genome_trs.bed |sort -k1,1 -k4,4
    n> cds_with_trs.gff
bedtools merge -i cds_with_trs.gff > cds_with_trs
    .bed
bedtools intersect -a known_genes.gff -b
    cds_with_trs.bed > known_genes_cds_trs.gff
bedtools merge -i cds_with_trs.gff > cds_with_trs
    .bed
bedtools intersect -a known_genes.gff -b
    cds_with_trs.bed > known_genes_cds_trs.gff
bedtools intersect -a cds_with_trs.bed -b
    pacbio_indels.bed >
    cds_with_trs_pacbio_indels.bed
bedtools intersect -a cds_with_trs.bed -b
    ilm_indels_gt_dp5.bed >
    cds_with_trs_ilm_indels_gt_dp5.bed
cat cds_with_trs_ilm_indels_gt_dp5.bed
    cds_with_trs_pacbio_indels.bed | sort -k1,1 -
    k2,2n > cds_with_trs_ilm_pb_indels_gt_dp5.bed
bedtools merge -i
    cds_with_trs_ilm_pb_indels_gt_dp5.bed >
    cds_with_trs_ilm_pb_indels_merged_gt_dp5.bed
bedtools intersect -a known_genes.gff -b
    cds_with_trs_ilm_pb_indels_gt_dp5.bed >
    cds_known_genes_intersect_trs_indel_ilm_pb_gt_dp5
    .gff
```

bedtools version 2.24.0 [112] was used to find the intersection between the coding sequence of genes with similarity to proteins from SwissProt/UniProtKB and TRs from Phobos. The result from this was intersected with indels called by FreeBayes (Illumina reads) and PBHoney (PacBio reads).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and material
Accession numbers at ENA (http://www.ebi.ac.uk/ena) for the gadMor2 assembly: WGS: CAEA020000001-CAEA020024862 Scaffolds: LN836027-LN845721 Chromosomes: LN845748-LN845770
The four original assemblies and gadMor2: doi.org/10.6084/m9.figshare.3408247

Competing interests
Jenny M. Ekholm and Paul Peluso are full-time employees at Pacific Biosciences, a company developing single molecule sequencing technologies.

Funding
Norwegian Research Council project number 199806 to KSJ.

Authors contributions
OKT and AJN performed the genome assemblies. OKT and WBR annotated the genome assembly. OKT carried out the tandem repeat analysis. JRM, BPW and JK designed and wrote the assembly algorithms. JME, PP, AT-K and MS performed library creation and sequencing. HG and SL created the linkage map. RBE contributed samples for sequencing and evaluated assemblies. OKT, BS, SJ, BPW, KSJ and AJN wrote the paper with help from all the authors. All authors read and approved the final manuscript

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; McGill University and Genome Quebec Innovation Centre, Canada and Pacific Biosciences of California Inc., Menlo Park. We are grateful for the willingness of Helle Tessand Baalsrud, Martin Malmstrøm and Monica H. Solbakken to investigate their favorite genes as a validation in different draft assemblies. Geir Kjetil Sandve provided essential help with implementing different algorithms for assembly reconciliation. We thank Jonas Korlach for his contribution to strategic aspects of the project and commenting on earlier versions of the manuscript.

Author information
All scripts are available at https://github.com/uio-cels/cod2_scripts and https://github.com/uio-cels/Repeats.

Author details
[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo., Oslo, Norway. [2]Department of Natural Sciences, University of Agder, Kristiansand, Norway. [3]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, NO-1432, Ås, Norway. [4]J. Craig Venter Institute, 9704 Medical Center Drive, 20850, Rockville, MD, USA. [5]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [6]Yale School of Medicine, Yale University, 06520, New Haven, CT, USA. [7]Pacific Biosciences, Menlo Park, CA, USA. [8]Institute of Marine Research, P.O. Box 1870, Nordnes, NO-5817, Bergen, Norway. [9]Biomedical Informatics Research Group, Department of informatics, University of Oslo., Oslo, Norway.

References
1. Ekblom, R., Wolf, J.B.W.: A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. (2014)
2. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., *et al.*: The sequence and de novo assembly of the giant panda genome. Nature **463**(7279), 311–317 (2010)
3. Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., Blomberg, L.A., *et al.*: Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. PLoS Biol. **8**(9), 1000475 (2010)
4. Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., *et al.*: The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. **43**(2), 109–116 (2011)

5. Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., et al.: The genome sequence of Atlantic cod reveals a unique immune system. Nature, 1–4 (2011)
6. IWGSC, T.I.W.G.S.C., Mayer, K.F.X., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., *et al.*: A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science **345**(6194), 1251788 (2014)
7. Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., *et al.*: The Norway spruce genome sequence and conifer genome evolution. Nature **497**(7451), 579–584 (2013)
8. Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., *et al.*: The oyster genome reveals stress adaptation and complexity of shell formation. Nature **490**(7418), 49–54 (2012)
9. Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., et al.: The Atlantic salmon genome provides insights into rediploidization. Nature (2016)
10. Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium,, *et al.*: Ancient hybridizations among the ancestral genomes of bread wheat. Science **345**(6194), 1250092 (2014)
11. Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., *et al.*: Comparative genomics reveals insights into avian genome evolution and adaptation. Science **346**(6215), 1311–1320 (2014)
12. Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., *et al.*: Whole-genome analyses resolve early branches in the tree of life of modern birds. Science **346**(6215), 1320–1331 (2014)
13. Alkan, C., Sajjadian, S., Eichler, E.E.: Limitations of next-generation genome sequence assembly. Nat Methods. **8**(1), 61–65 (2011)
14. Chaisson, M.J.P., Wilson, R.K., Eichler, E.E.: Genetic variation and the de novo assembly of human genomes. Nature Rev Genet. **16**(11), 627–640 (2015)
15. Treangen, T.J., Salzberg, S.L.: Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Rev Genet. **13**(1), 36–46 (2012)
16. Miller, J.R., Koren, S., Sutton, G.G.: Assembly algorithms for next-generation sequencing data. Genomics **95**(6), 315–327 (2010)
17. Chalopin, D., Naville, M., Plard, F., Galiana, D., Volff, J.-N.: Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. (2015)
18. Mayer, C., Leese, F., Tollrian, R.: Genome-wide analysis of tandem repeats in *Daphnia pulex* - a comparative approach. BMC Genom. **11**, 277 (2010)
19. Gemayel, R., Vinces, M.D., Legendre, M., Verstrepen, K.J.: Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. **44**(1), 445–477 (2010)
20. Roberts, R.J., Carneiro, M.O., Schatz, M.C.: The advantages of SMRT sequencing. Genome Biol. **14**(6), 405 (2013)
21. Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., *et al.*: Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. **30**(7), 693–700 (2012)
22. Conte, M.A., Kocher, T.D.: An improved genome reference for the African cichlid, *Metriaclima zebra*. BMC Genom. **16**(1), 1 (2015)
23. Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., *et al.*: Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. **12**(8), 780–786 (2015)
24. Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., Phillippy, A.M.: Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. **33**(6), 623–630 (2015)
25. Vij, S., Kuhl, H., Kuznetsova, I.S., Komissarov, A., Yurchenko, A.A., van Heusden, P., *et al.*: Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. PLoS Genet. **12**(4), 1005954 (2016)
26. Braasch, I., Peterson, S.M., Desvignes, T., McCluskey, B.M., Batzel, P., Postlethwait, J.H.: A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. J Exp Zool B **324**(4), 316–341 (2015)
27. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., *et al.*: Ensembl 2014. Nucleic Acids Res. **42**(D1), 749–755 (2013)
28. Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., *et al.*: GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. **22**(3), 557–567 (2012)
29. Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., *et al.*: Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. **21**(12), 2224–2241 (2011)
30. Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., *et al.*: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience **2**(1), 10 (2013)
31. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., *et al.*: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA **108**(4), 1513–1518 (2011)
32. Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., *et al.*: Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**(24), 2818–2824 (2008)
33. Walker, B.J., 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 **9**(11), 112963 (2014)
34. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., *et al.*: Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLOS ONE **7**(11), 47768 (2012)
35. Vezzi, F., Narzisi, G., Mishra, B.: Reevaluating assembly evaluations with feature response curves: GAGE and Assemblathons. PLOS ONE **7**(12), 52210 (2012)
36. Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D.: REAPR: a universal tool for genome assembly evaluation. Genome Biol. **14**(5), 47 (2013)
37. Ryan, J.F.: Baa.pl: a tool to evaluate de novo genome assemblies with RNA transcripts. arXiv.org (2013)
38. Parra, G., Bradnam, K., Korf, I.F.: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**(9), 1061–1067 (2007)
39. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19), 3210–3212 (2015)
40. Yao, G., Ye, L., Gao, H., Minx, P., Warren, W.C., Weinstock, G.M.: Graph accordance of next-generation sequence assemblies. Bioinformatics **28**(1), 13–16 (2011)
41. Wences, A.H., Schatz, M.C.: Metassembler: merging and optimizing de novo genome assemblies. Genome Biol. **16**(1), 1 (2015)
42. Angiuoli, S.V., Salzberg, S.L.: Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics **27**(3), 334–342 (2011)
43. Simpson, J.T., Durbin, R.: Efficient construction of an assembly string graph using the FM-index. Bioinformatics **26**(12), 367–373 (2010)
44. Simpson, J.T.: Exploring genome characteristics and sequence quality without a reference. Bioinformatics **30**(9), 023–1235 (2014)
45. Hardie, D.C., Hebert, P.: The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome **46**(4), 683–706 (2003)
46. Hardie, D.C., Hebert, P.D.: Genome-size evolution in fishes. Can J Fish Aquat Sci. **61**(9), 1636–1646 (2004)
47. Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., *et al.*: MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. **164**(2), 513–524 (2014)
48. Holt, C., Yandell, M.: MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. **12**(1), 491 (2011)
49. Tørresen, O.K., Samy, J.K.A., Våge, D.I., Nederbragt, A.J.: A Genome Browser for the Atlantic Cod Genome Version 2. `http://www.mn.uio.no/cees/english/research/about/infrastructure/genome-browser/`
50. Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., *et al.*: The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**(7392), 55–61 (2012)
51. 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. **6**, 21902 (2016)

52. 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 **5**, 12081 (2016)
53. Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., *et al.*: Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. Genome Res. **15**(8), 1127–1135 (2005)
54. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.org (2013)
55. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing. arXiv.org (2012)
56. Chaisson, M.J.P., Tesler, G.: Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinform. **13**(1), 238 (2012)
57. English, A.C., Salerno, W.J., Reid, J.G.: PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping. BMC Bioinform. **15**(1), 180 (2014)
58. Smit, A., Hubley, R.: RepeatModeler Open-1.0. `http://www.repeatmasker.org`
59. Ellinghaus, D., Kurtz, S., Willhoeft, U.: *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. **9**(1), 1 (2008)
60. Steinbiss, S., Willhoeft, U., Gremme, G., Kurtz, S.: Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. Nucleic Acids Res. **37**(21), 7002–7013 (2009)
61. Haas, B.J.: TransposonPSI. `http://transposonpsi.sourceforge.net`
62. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J.: Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. **110**(1-4), 462–467 (2005)
63. Jiang, Q., Li, Q., Yu, H., Kong, L.: Genome-wide analysis of simple sequence repeats in marine animals—a comparative approach. Mar Biotechnol. (New York, N.Y.) **16**(5), 604–619 (2014)
64. Gymrek, M., Golan, D., Rosset, S., Erlich, Y.: lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. **22**(6), 1154–1162 (2012)
65. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**(2), 573–580 (1999)
66. Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, Mittelman, D., Erlich, Y.: The landscape of human STR variation. Genome Res. **24**(11), 1894–1904 (2014)
67. Parra, G., Bradnam, K., Ning, Z., Keane, T., Korf, I.F.: Assessing the gene space in draft genomes. Nucleic Acids Res. **37**(1), 289–297 (2009)
68. Li, H.: Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics **30**(20), 2843–2851 (2014)
69. Jiang, Y., Turinsky, A.L., Brudno, M.: The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Res. **43**(15), 677–7228 (2015)
70. Star, B., Hansen, M.H., Skage, M., Bradbury, I.R., Godiksen, J.A., Kjesbu, O.S., *et al.*: Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. Sci Technol Archaeol Res. **2**(1), 36–45 (2016)
71. Ellegren, H.: Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16**(12), 551–558 (2000)
72. Ellegren, H.: Microsatellites: simple sequences with complex evolution. Nature Rev Genet. **5**(6), 435–445 (2004)
73. Firtina, C., Alkan, C.: On genomic repeats and reproducibility. Bioinformatics (2016)
74. Usdin, K.: The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. **18**(7), 1011–1019 (2008)
75. Hammock, E.A.D., Young, L.J.: Microsatellite instability generates diversity in brain and sociobehavioral traits. Science **308**(5728), 1630–1634 (2005)
76. Fondon III, J.W., Garner, H.R.: Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA **101**(52), 18058–18063 (2004)
77. Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A., *et al.*: Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLOS ONE **8**(2), 54710 (2013)
78. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., Verstrepen, K.J.: Unstable tandem repeats in promoters confer transcriptional evolvability. Science **324**(5931), 1213–1216 (2009)
79. 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. (2015)
80. Sonay, T.B., Carvalho, T., Robinson, M., Greminger, M., Krutzen, M., Comas, D., *et al.*: Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res. **25**(11), 190868–1151599 (2015)
81. Ohadi, M., Valipour, E., Ghadimi Haddadan, S., Namdar Aligoodarzi, P., Bagheri, A., Kowsari, A., *et al.*: Core promoter short tandem repeats as evolutionary switch codes for primate speciation. Am J Primatol. **77**(1), 34–43 (2015)
82. Verstrepen, K.J., Jansen, A., Lewitter, F., Fink, G.R.: Intragenic tandem repeats generate functional variability. Nat Genet. **37**(9), 986–990 (2005)
83. Lindqvist, C., Laakkonen, L., Albert, V.A.: Polyglutamine variation in a flowering time protein correlates with island age in a Hawaiian plant radiation. BMC Evol Biol. **7**(1), 1 (2007)
84. Ottersen, G., Bogstad, B., Yaragina, N.A., Stige, L.C., Vikebo, F.B., Dalpadado, P.: A review of early life history dynamics of Barents Sea cod (*Gadus morhua*). ICES J Mar Sci., 2064–2087 (2014)
85. Poulsen, N.A.A., Nielsen, E.E., Schierup, M.H., Loeschcke, V., Grønkjaer, P.: Long-term stability and effective population size in North Sea and Baltic Sea cod (*Gadus morhua*). Mol Ecol. **15**(2), 321–331 (2006)
86. Kent, W.J.: BLAT–the BLAST-like alignment tool. Genome Res. **12**(4), 656–664 (2002)
87. Magoc, T., Salzberg, S.L.: FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27**(21), 2957–2963 (2011)
88. Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B.P., *et al.*: The bonobo genome compared with the chimpanzee and human genomes. Nature **486**(7404), 527–531 (2012)
89. Li, H.: Toolkit for Processing Sequences in FASTA/Q Formats. `https://github.com/lh3/seqtk`
90. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., *et al.*: Versatile and open software for comparing large genomes. Genome Biol. **5**(2), 12 (2004)
91. Hubert, S., Higgins, B., Borza, T., Bowman, S.: Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua). BMC Genom. **11**(1), 191 (2010)
92. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al.*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. **29**(7), 644–652 (2011)
93. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al.*: Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**(7057), 376–380 (2005)
94. Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., *et al.*: Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLOS ONE **10**(7), 0132628 (2015)
95. Penglase, S., Edvardsen, R.B., Furmanek, T., Rønnestad, I., Karlsen, Ø., van der Meeren, T., *et al.*: Diet affects the redox system in developing Atlantic cod (*Gadus morhua*) larvae. Redox Biol. **5**, 308–318 (2015)
96. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10–12 (2011)
97. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., *et al.*: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. **8**(8), 1494–1512 (2013)
98. Kleppe, L., Edvardsen, R.B., Kuhl, H., Malde, K., Furmanek, T., Drivenes, Ø., *et al.*: Maternal 3'UTRs: from egg to onset of zygotic transcription in Atlantic cod. BMC Genom. **13**(1), 443 (2012)

99. Llorens, C., Muñoz-Pomer, A., Futami, R.: The GyDB collection of viral and mobile genetic element models. Biotechvana Bioinf. (2009)
100. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5), 1792–1797 (2004)
101. Kennedy, R.C., Unger, M.F., Christley, S., Collins, F.H., Madey, G.R.: An automated homology-based approach for identifying transposable elements. BMC Bioinform. **12**(1), 1 (2011)
102. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28**(23), 3150–3152 (2012)
103. UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res. **43**(Database issue), 204–12 (2015)
104. Cantarel, B.L., Korf, I.F., Robb, S.M.C., Parra, G., Ross, E., Moore, B., *et al.*: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**(1), 188–196 (2008)
105. Kumar, S.: How to Predict Genes Using a Two-pass (iterative) MAKER2 Workflow. `https://github.com/sujaikumar/assemblage/blob/master/README-annotation.md`
106. Campbell, M.S., Holt, C., Moore, B., Yandell, M.: Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics **48**, 4–11141139 (2014)
107. Korf, I.F.: Gene finding in novel genomes. BMC Bioinform. **5**(1), 59 (2004)
108. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., Borodovsky, M.: Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**(20), 6494–6506 (2005)
109. Stanke, M., Waack, S.: Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**(suppl 2), 215–225 (2003)
110. Stanke, M., Diekhans, M., Baertsch, R., Haussler, D.: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**(5), 637–644 (2008)
111. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., et al.: InterProScan 5: genome-scale protein function classification. Bioinformatics (2014)
112. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)

Additional Files

Additional file 1 — Supplementary Information
Supplementary figures and tables.