Chromosome-level genome assembly of burbot (Lota lota) provides insights into the evolutionary adaptations in freshwater

Zhiqiang Han[1], Manhong Liu[2], Qi Liu[3], Hao Zhai[2], shijun xiao[3], and Tianxiang Gao[4]

[1]Affiliation not available
[2]Northeast Forestry University
[3]Wuhan Gooalgene Technology Company
[4]Zhejiang Ocean University

October 8, 2020

Abstract

The burbot (Lota lota) is the only member of the cod family (Gadidae) that is adapted solely to freshwater. This species shows the widest longitudinal range of freshwater fish in the world. The burbot is a good model for studies on adaptive genome evolution from marine to freshwater environment. However, no high-quality reference genome has been released. Here, the first chromosome-level genome of the burbot was constructed using PacBio long sequencing and Hi-C technology. A total of 95.24 Gb polished PacBio sequences were generated, and the preliminary genome assembly was 575.83 Mb in size with a contig N50 size of 2.15 Mb. The assembled sequences were anchored to 22 pseudo-chromosomes by using the Hi-C data. The final assembled genome after Hi-C correction was 575.92 Mb, with a contig N50 of 2.01 Mb and a scaffold N50 of 22.10 Mb. A total of 22,067 protein-coding genes were predicted, 94.82% of which were functionally annotated. Phylogenetic analyses indicated that burbot diverged with the Atlantic cod about 44.4 million years ago. In addition, 377 putative genes that appear to be under positive selection in burbot were identified. These positively selected genes might adapt to the freshwater environment. These genome data provide an invaluable resource for the ecological and evolutionary study of the order Gadiformes.

Introduction

The order Gadiformes includes some of the most important commercial fish (e.g., cod, hake, and haddock) in the world and accounts for approximately 18% of the world's total marine fish catch (FAO, 2004). Gadiform fish inhabit cold waters in every high-latitude ocean from deep-sea benthic habitats to coastal waters. Only two species in this order are known in freshwater habitats (Nelson, 2006). However, to date, only one high-quality genome sequence of the Gadiformes species, i.e., the Atlantic cod (*Gadus morhua*) (Star et al., 2011), is available, and this limitation significantly hinders the taxonomical, evolutionary, and biological studies of the order Gadiformes.

The burbot *Lota lota* is the only member of the cod family (Gadidae) that is adapted solely to freshwater (Schaefer et al., 2016). This fish has a wide holarctic distribution, showing the widest longitudinal range of freshwater fish in the world. The burbot is distributed in nearly all suitable freshwater basins of North America, Europe, and north Asia (Lehtonen, 1998). Although this fish thrives in freshwater, *L. lota* has retained many characteristics of its marine ancestors (Blabolil et al., 2018), such as preference for cold water, spawning at low temperatures, high fecundity, and a pelagic larval stage. This species spawns during winter or early spring, typically when the water is still ice-covered, and the water temperatures are between 1 °C and 4 °C (Bergersen et al., 1993). Spawning occurs on fine to gravel substrate in shallow bays or groyne fields in water depths of 0.3 m to 3.0 m (Fredrich & Arzbach, 2002; Eick et al., 2013).

The burbot is apparently an excellent "indicator" species. This species is vulnerable to many environmental changes, in particular, warming water temperatures and pollution (Stapanian et al., 2010). The burbot in marginal habitats may serve as an early indicator of the impacts of climate change on cold-water fish species (Stapanian et al., 2010). However, stocks of the burbot have severely declined in number and distribution during the past century. Many populations are threatened, have been extirpated, or are otherwise in need of conservation measures (Maitland & Lyle, 1990). For example, in Finland, burbot populations have declined or have been destroyed completely in 16% of the lakes (Tammi et al., 1999). A series of threats, including pollution, habitat fragmentation, exploitation, and invasive species, have caused the decline or extirpation of many burbot populations (Stapanian et al., 2010). Genomics resources will support the conservation studies of burbot. Given the widest holarctic distribution of this species, the burbot may undergo some degree of local adaptation, which can be resolved with genome-wide high-quality SNPs. However, the available genetic information for this fish remains scarce. At present, only limited genetic studies have been conducted on the microsatellite loci isolation and population structure of the burbot (Houdt et al., 2005; Sanetra, 2005). Thus, sequencing the genome of the burbot is essential. This process may help to reveal insights into the evolutionary history of the burbot and the role of environmental changes in shaping the genome evolution from marine to freshwater.

The burbot,the only freshwater species in the cod family (Gadidae),represents a classical transition from marine to freshwater. Fossil evidence suggests that the *Lota* genus has already inhabited European rivers in the early Pliocene (Houdt et al., 2005). This phenomenon indicates that the burbot left the ocean and migrated to the freshwater. The transition from an oceanic to a freshwater habitat provides an opportunity for drastic environmental changes in the ecology, morphology, and behavior of fish. This transition should select numerous functional genes. Marine to freshwater transition events rarely occurs (Finnegan, 2017), which is likely due to physiological and ecological barriers associated with changing environmental conditions. These factors include lower salinity, relatively high levels of UV radiation, dramatic fluctuations in freshwater temperature, and competition from primary freshwater fish lineages. Despite these challenges, several freshwater fish from marine-derived lineages have completed this transition from ocean to freshwater and invaded successfully freshwater habitats. Concerted effort, such as the convergence of their morphological and physiological characters, has been reported to freshwater adaptations (Jara, 1988). The genomic changes underlying a convergent evolution may be reproducible to some extent, and convergent phenotypic traits may commonly arise from the same genetic changes. Physiological convergence is strong in freshwater fish of marine-derived lineages and provides a practical way to identify freshwater adaptations.

In this study, a chromosome-level genome assembly of burbot was constructed by combining short reads, PacBio long reads, and Hi-C sequencing data. The assembly was used to identify the genetic signatures of evolution related to freshwater adaptation in burbot and Perciformes by comparative genomics of 13 distantly related species, including three freshwater Percomorpha species. This study will provide a genomic resource to further address the key evolutionary process of freshwater adaptation for marine-originated species.

Materials and methods

Sample and DNA extraction

A single female fish (~800 g) was collected in November 2019 from Heilong River, the northeastern part of China. Muscle, eye, gonad, gill, liver, and spleen tissues were collected and stored in liquid nitrogen until DNA and RNA extraction. Muscle tissues were used for DNA sequencing for genome assembly, while all tissues were used for transcriptome sequencing. Genomic DNA from the white muscle tissue was extracted using the standard phenol/chloroform extraction method to construct the DNA sequencing library. The integrity and concentration of the genomic DNA molecules were checked using 1% agarose gel electrophoresis and Pultton DNA/Protein Analyzer (Plextech, USA).

Library construction and genome sequencing

The Illumina NovaSeq-6000 and PacBio Sequel II platforms were applied for genomic sequencing to generate short and long genomic reads, respectively. Illumina sequencing libraries were prepared to estimate the genome size, correct the genome assembly, and evaluate assemblies. A paired-end library was constructed with an insert size of 300 bp according to the Illumina standard protocol. After discarding reads with low-quality bases (reads with more than 10% N bases or low-quality bases[?]5), adapter sequences, and duplicated sequences, the clean reads were used for subsequent analysis.

For long-read sequencing, we constructed an SMRTbell library with a fragment size of 20 Kb by using the SMRTBell template preparation kit 1.0 (PacBio, USA) by following the manufacturer's protocol. The library was sequenced with the PacBio Sequel II system, and data from one SMRT cell were generated.

Genome size estimation and genome assembly

The Kmer-based method of the Illumina short read data was used to analyze the genome survey to estimate the genome size, heterozygosity, and repeat content of the burbot genome (Liu et al., 2013). The NextDenovo package (https://github.com/Nextomics/NextDenovo) was performed to assemble the burbot genome with PacBio long reads by using the following parameters: parallel_jobs, 300; seed_cutoff, 15,000; pa_correction, 320; and random_round, 100. To correct the random sequencing errors in the NextDenovo output, two steps of genome sequence polishing were applied. The arrow was used to polish the genome by using the long sequencing data (Chin et al., 2013), and two rounds of polishing using Illumina short reads were then applied with Pilon (Walker et al., 2014).

Hi-C analysis and chromosome assembly

To obtain a chromosome-scale genome assembly, the Hi-C library for sequencing was constructed. The muscle tissue of the burbot was used to prepare the library, according to Rao et al. (2014). High-quality Hi-C fragment libraries were sequenced for the Illumina NovaSeq-6000 platform. The sequencing reads were mapped to the polished burbot genome with Bowtie 1.2.22. The two read ends were independently aligned to the genome, and only the read pairs with both ends uniquely aligned to the genome were selected. Lachesis (Burton et al., 2013) with default parameters was then applied to perform the chromosomal-level genome assembly by using the corrected contigs and valid Hi-C reads. The ggplot2 in the R package was applied to generate a genome-wide Hi-C heatmap to evaluate the quality of the chromosomal-level genome assembly.

Two methods were performed to assess the completeness and accuracy of the genome assembly. First, the Illumina and PacBio reads were aligned to the burbot assembly genome by using BWA-MEM (version 0.7.10-r789) (Li & Durbin, 2009) and BLASR (Mark et al., 2012), respectively. Second, the completeness of the genome assembly was evaluated by using BUSCO (version 2.0) (Simao et al., 2015) to search the genome against the Actinopterygii database, which consisted of 4,584 single copy orthologs.

Repeat annotation, gene prediction and functional annotation

Repeat elements were annotated in the burbot genome before the gene prediction. Tandem Repeat Finder (Benson, 1999) and LTR_FINDER (Zhao et al., 2007) were applied for the *ab initio* prediction of repeat elements in this genome. RepeatMasker and RepeatProteinMask (*http://www.repeatmasker.org*) were executed to search the genome sequences for known repeat elements, with the genome sequences used as queries against the Repbase database (Jurka et al., 2005). The ribosomal RNA (rRNA) and microRNA genes were predicted using the Infernal v.1.1 software based on the Rfam and miRBase databases, respectively (Griffiths-Jones et al., 2005). Transfer RNA (tRNA) genes were identified using the tRNAscan-SE v. 1.3.1 software (Lowe & Eddy, 1997).

Based on the repeat-masked genome, three strategies based on *abinitio* , homologs, and RNA-sequencing were applied to predict the protein-coding genes in the burbot genome assembly. *Ab initio*gene prediction was performed using Augustus (v2.7) (Stanke et al., 2006) and GenScan (Burge & Karlin, 1997) with default settings. For the homology-based prediction, protein sequences of *Gadus morhua* ,*Acanthochromis polyacanthus* , *Oryzias latipes* ,*Amphiprion ocellaris* , *Anabas testudineus* ,*Astatotilapia calliptera* , *Astyanax mexicanus* ,*Austrofundulus limnaeus* , *Lepisosteus oculatus* , and*Notothenia coriiceps* were downloaded from the NCBI database and aligned to the burbot genome by using tBLASTn (E-value [?] 1e-5). The homologous genome sequences were then aligned against the matching proteins by using GeneWise (v2.4.0) (Doerks et al., 2002) for accurately spliced alignments. Transcriptomic data (2x150 bp, NovaSeq-6000 platform) generated from a mixture of six tissues, including the muscle, eye, gonad, gill, liver, and spleen, were aligned to the assembled genome sequences by using HISAT2 (v2.0.10) (Pertea et al., 2016), and the putative transcript structures were detected using gene structure, which was predicted by Cufflinks (Ghosh et al., 2016). All gene models were merged, and redundancy was removed by MAKER (Cantarel et al., 2008) and HiCESAP.

The NCBI nonredundant protein (NR), InterPro, the SwissProt, TrEMBL (Boeckmann et al., 2003), eukaryotic orthologous groups of proteins (KOG) (Tatusov et al., 2003), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) databases with an E-value threshold of 1e–5 were used for the functional annotation of the protein-coding genes by using BLASTX and the BLASTN utility (Lobo, 2008). Functional ontology and pathway information from the Gene Ontology (GO) database was assigned to the genes by using Blast2GO (Conesa et al., 2005).

Comparative genomic analyses and testing for genomic selection

The protein sequences of 12 species of teleost fish (Supporting Information Table S1) were downloaded from Ensembl (release version 100). Only the longest transcript was selected for each gene locus with alternative splicing variants. Orthologous groups were constructed by ORTHOMCL (14) v2.0.9 by using the default settings based on the filtered BLASTP results. The single-copy orthologous genes shared by all 13 species were further aligned using MUSCLE (version 3.8.31) (Edgar, 2004) and concatenated to construct a phylogenetic tree with RaxML (Stamatakis, 2014). The divergence time among species was estimated by the r8s (Sanderson, 2003). Divergence times of*Larimichthys crocea-Notothenia coriiceps* (88–114), *L. crocea-Gambusia affinis* (96.9–150.9), *Clupea harengus-G. affinis* (149.85–165.2) from the TimeTree database (Kumar et al., 2017) were used as the calibration times. Gene family expansion and contraction analyses were performed using CAFE 3.1 (De Bie et al., 2006) with the estimated phylogenetic tree information. P value < 0.05 was used to indicate significantly changed gene families. The expanded and contraction gene families in burbot, the shared contraction gene families of three freshwater species (burbot,*Monopterus albus,* and *G. affinis*) in GO terms, and KEGG pathways were enriched, and the Benjamini and Hochberg FDR correction was applied. Significantly overrepresented GO terms and KEGG pathways were identified with corrected P values [?]0.05.

To construct multiple sequence alignments among the ortholog genes, the CODEML program in PAML 4.5 was used to estimate the dN/dS ratio (ω) (Yang, 2007). Two different branch-site likelihood ratio tests were applied to find genes under positive selection. First, the species-specific positively selected genes (PSGs) in burbot were identified with burbot as foreground species and other species, excluding zebrafish *Danio rerio* , *Clupea harengus* and*Esox lucius* as background species. Second, three freshwater species (i.e., burbot, *M. albus,* and *G. affinis*) were selected as foreground species and other species, excluding zebrafish*Danio rerio* , *Clupea harengus* and *Esox lucius* , as background species. GO and KEGG categories were assigned to orthologous groups according to the zebrafish genome reference for the functional enrichment analyses.

3 Results and discussion

3.1Genome size estimation and initial characterization of the genome

A total of 79 Gb of Illumina data werer generated from the Illumina 150 insert-size library, representing 143.64-fold coverage of the burbot genome (Table 1, Supplementary Table S2). The total number of fk -mers was 62,863,455,719, with a k -mers peak at a depth of 112 (Supplementary Figure S1 and Table S3). The genome size was estimated to be ˜550 Mb with heterozygosity of 0.57% and repeat content of 36.60% (Supplementary Table S3).

3.2 Genome assembly and completeness of the assembled genome

The PacBio Sequel II platform generated 95.24 Gb high-quality data from the long-read library, covering 173.16-fold of the genome assembly (Table 1, Supplementary Table S4). These data were assembled using NextDenovo followed by racon and pilon polishing, which produced a 575.83 Mb genome assembly with a contig N50 of 2.15 Mb (Table 2). The length of this assembly was consistent with the genome size estimated by k-mer analysis.

The Illumina reads and PacBio long reads were aligned to the burbot assembly to evaluate the quality of the initial genome assembly. The results showed that 99.23 % of the Illumina reads and 97.55% of the PacBio long reads were successfully mapped to the assembled genome (Supplementary Table S5 and S6). The BUSCO analysis showed that 94.67% (4344/4584) of the complete BUSCO were found in the genome assembly (Table 3), including 91.93% of the complete and single copy and 2.84% duplicated genes.

The contigs in the draft assembly were then anchored and oriented into a chromosomal-scale assembly by using the Hi-C scaffolding approach. The Hi-C library generated 69.51 Gb (126.38×) clean data (Table 1, Supplementary Table S7). With the use of LACHESIS, 88.66% of the assembled sequences were anchored to 22 pseudo-chromosomes, with chromosome lengths ranging from 15.18 Mb to 51.8 Mb (Table 4). Based on the heatmap, the 22 pseudochromosomes could be distinguished easily and the interaction signal strength around the diagonal was considerably strong, which indicated a high quality of this genome assembly (Figure 2). The final assembled genome after Hi-C correction was 575.92 Mb, with a contig N50 of 2.01 Mb and a scaffold N50 of 22.10 Mb (Table 5).

3.3 Repeat annotation, gene prediction and gene annotation

A total of 384.29 Mb of repeat sequences were detected, accounting for 66.74% of the assembly genome (Table 6). This repeat content was obviously larger than the value (36.60%) obtained from the k-mer analysis. The repetitive sequences mainly consisted of the DNA transposable element (289.32 Mb; 50.24% of the assembly), long terminal repeats (66.95 Mb; 11.63%), and long interspersed elements in 30.96Mb (5.38%) (Table 7).

A total of 21,664 protein-coding genes were predicted by the combination of strategies based on *ab initio* , homologs, and RNAseq. The average values of the gene length, exon length, and average intron length were 14,606, 292.38, and 1,223 bp, respectively (Table 8). The statistics of the predicted gene models were compared to other ten teleost species,including :*Acanthochromis polyacanthus,Oryzias latipes,Amphiprion ocellaris,Anabas testudineus,Astatotilapia calliptera,Astyanax mexicanus,Austrofundulus limnaeus,Gadus morhua,Lepisosteus oculatus,Notothenia coriiceps* , showing similar distribution patterns in mRNA length, CDS length, exon length, intron length and exon number (Supplementary Figure S2). The summary of genome characteristics of burbot was shown in Figure 3. A total of 20658 predicted genes (95.36%) were successfully annotated by alignment to the nucleotide, protein, and annotation databases InterPro, NR, Swissprot, TrEMBL, KOG, GO, and KEGG (Table 9). A total of 6390 tRNAs, 300 rRNAs, and 519 microRNAs were identified by noncoding RNA prediction (Supplementary Table S8).

3.4 **Comparative genomics and the mechanism of adaption to freshwater**

A total of 19,998 gene families and 2,650 single-copy orthologous genes were identified using the genomes and genes of 13 selected teleosts. In addition, 21,664 genes of burbot could be clustered into 14,504 gene families, including 132 unique gene families (Supplementary Table S9). Based on the single-copy orthologous genes, the ML phylogenetic tree was constructed and showed that burbot and Atlantic cod were clustered together, and the divergence time between two cod species was ˜44.4 Mya (Figure 4). The divergence time was consistent with the estimated time by Hughes et al. (2018). The burbot genome displayed 639 expanded and 1564 contracted gene families compared with the common ancestor of burbot and Atlantic cod (Figure 4). The expanded gene families of burbot were significantly enriched in 73 GO terms and 34 KEGG pathways, mainly including DNA integration (GO:0015074, corrected *P* value =0.00E+00), DNA metabolism process (GO:0006259, corrected *P* value =2.05E-06), apoptosis process (GO:0006915, corrected *P* value =5.22E-05), zinc ion binding (GO:0008270, corrected *P* value =2.02E-96), transition metal ion binding (GO:0046914, corrected *P* value =1.19E-91), natural killer cell-mediated cytotoxicity (ko04650, corrected *P* value =4.200299E-20), and hematopoietic cell lineage (ko04640, corrected*P* value=3.04E-18) that were associated with cell damage repair, ion binding, and immune system (Supplementary Tables S10 and S11). Conversely, the burbot clearly showed contracted gene families in homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156, corrected *P* value =2.69E-29), cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742, corrected *P* value =2.69E-29), membrane (GO:0016020, corrected *P* value=1.18E-10) GO terms, amino sugar and nucleotide sugar metabolism (ko00520, corrected*P* value=1.52E-04), and NOD-like receptor signaling (ko04621, corrected *P* value=2.41E-02) pathways (Supplementary Tables S12 and S13).

Notably, three freshwater species shared no expanded gene families and two contracted gene families associated with cell adhesion (GO:0007155: corrected *P* value=0.00E+00) and membrane (GO:0016020, corrected*P* value =0.00E+00) (Supplementary Table S14). These functions are critical for adjusting the ion concentrations inside and outside the cell. However, no enriched KEGG pathway was found for the contracted gene families. Such gene families may reflect the reduced functional requirements of a stable ionic environment in freshwater for cell membrane permeability. These findings are consistent with the different components of omega-3 fatty acids between marine and freshwater fish (Taşbozan & Gökçe, 2017). Marine fish have higher levels of omega-3 fatty acids than freshwater species. Compared with the omega-6 fatty acids, omega-3 fatty acids help improve cell membrane fluidity and provide osmoregulatory capabilities.

To identify the genes evolving under positive selection for freshwater adaptation, two different likelihood ratio tests (branch-site model) were performed. A total of 377 genes were identified as PSGs in the burbot genome (Supplementary Table S15). The burbot PSGs were functionally enriched in the organic cyclic compound metabolic process (GO:1901360, corrected *P* value =1.83E-02), cellular nitrogen compound metabolic process (GO:0034641, corrected *P* value =4.11E-03), RNA metabolic process (GO:0016070, corrected *P* value =4.13E-03), and nucleic acid metabolic process (GO:0090304, corrected*P* value =6.16E-03) (Supplementary Table S16). Additionally, 38 PSGs were detected with three freshwater lineages (burbot, *M. albus* and *G. affinis*) as foreground branch (Supplementary Table S17). Four PSGs (*stk33* , *ino80e* , *nabp1a* and*znf385a*) were related to DNA damage repair. Genes *stk33 and nabp1*participate in the mitotic DNA damage checkpoint. *znf385a* is located upstream in the p53 activating pathway. *znf385a* interacts with p53/TP53 and promotes DNA damage-induced cell cycle arrest (Das et al., 2007). Protein ino80e is a component of the chromatin remodeling INO80 complex and contributes to the DNA double-strand break repair (Yao et al., 2008).

The exposure of freshwater fish to UV radiation may cause DNA damage. The presence of a group of genes involved in DNA repair under positive selection was consistent with the high levels of exposure to UV radiation in freshwater environment compared with that in the ocean environment. This finding suggests that these genes had functionally convergent in three freshwater lineages.

The PSGs of freshwater lineages were enriched in folic acid transport (GO: 0015884, *slc19a1* , corrected *P* value =8.10E-05) GO terms, amino acid metabolism, replication, and repair pathways (Supplementary Tables S18 and S19). *slc19a1* has an important role in folate transmembrane transport. Low osmotic pressure has been previously shown to affect the efficiency of folic acid absorption in the intestine (Zhao et al.,2011). The positive selection on*slc19a1* may improve folic acid absorption for freshwater species. These data will serve as valuable resources for future evolution studies of burbot.

4. Conclusion

A chromosomal-scale genome assembly of the burbot was provided by integrating the Hi-C and PacBio long read sequencing data. The burbot is the only freshwater member of the cod family and represents the widest longitudinal range of freshwater fish in the world. The genome assembly and annotation supplied the second high-quality genome of the order Gadiformes and important genomic data for whole genome analysis to further investigate the evolution of burbot with other cod species. A series of candidate genes involved in freshwater adaptation were identified in these comparative genomics analyses. The results were beneficial in elucidating the evolution process in order Gadiformes under environment change. These data are also useful for diverse conservation applications, including identifying conservation units, assessing gene flow, detecting local adaptation of the populations and elucidating the evolutionary history of burbot.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2017YFA0604904) and Natural Science Foundation of Heilongjiang Province of China (LH2019C070).

Author contributions

T.X. G. and Z.Q.H. conceived and managed the project. M.H.L. and H.Z. collected the sequencing samples. Q.L. and X.S.J. extracted the DNA/RNA and performed the genome sequencing. Z.Q.H. analyzed the data. Z.Q.H. and T.X. G. wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of interests

No

Data Accessibility Statement

Raw sequencing data (PacBio, Illumina, Hi-C and RNA-seq data) for burbot genome has been deposited at the Sequence Read Archive (SRA) (SRR12549297, SRR12549430, SRR12550374 and SRR12577979). The assembled genome and annotations have been deposited at NCBI Assembly database with the GenBank accession PRJNA663985. In addition, the genome data is also openly available in figshare at*https://doi.org/10.6084/m9.figshare.12927203*(Liu, 2020).

References

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

Bergersen EP, Cook MF, Baldes RJ (1993) Winter movements of burbot (*Lota lota*) during an extreme drawdown in Bull Lake, Wyoming, USA. *Ecol Freshw Fish* , **2** , 141-145.

Blabolil P, Duras J, Jůza T, Kočvara L, Matěna J, Muška M, Říha M, Vejřík L, Holubová M, Peterka J (2018) Assessment of burbot *Lota lota* (L. 1758) population sustainability in central European reservoirs. *J Fish Biol* , **92** , 1545-1559.

Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin M, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* , **31** , 365-370.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* , **268** , 78.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* , **31** , 1119-1125.

Cantarel BL, Korf L, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* , **18** , 188-196.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* , **10** , 563.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* , **21** , 3674.

Das S, Raj L, Zhao B, Kimura Y, Bernstein A, Aaronson SA, Lee SW (2007) Hzf Determines cell survival upon genotoxic stress by modulating p53 transactivation. *Cell* , **130** , 624-637.

De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* ,**22** , 1269-1271.

Doerks T, Copley RR, Schultz J, Ponting CP, Bork P (2002) Systematic identification of novel protein domain families associated with nuclear functions. *Genome res* , **12** , 47-56.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* , **32** , 1792-97.

Eick D (2013) Habitat preferences of the burbot (*Lota lota*) from the River Elbe: An experimental approach. *J Appl Ichthyol* ,**29** , 541-548.

Finnegan D (2017) Convergence in Diet and Morphology in Marine and Freshwater Cottoid Fishes.

Fredrich F, Arzbach HH (2002) Migration and bank structure use of burbot, Lota lota in the River Elbe, Germany. *Zeitschrift für Fischkunde* , **(Suppl 1.)** , 159-178.

Ghosh S, Chan CKK (2016) Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods in Molecular Biology* , **1374** , 339.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Bateman A (2005) Rfam: Annotating Non-Coding RNAs in Complete Genomes. *Nucleic Acids Res* , **33** , D121-D124.

Houdt JKJV, Cleyn LD, Perretti A, Volckaert F (2005) A mitogenic view on the evolutionary history of the Holarctic freshwater gadoid, burbot (*Lota lota*). *Mol Ecol* , **14** , 2445-2457.

Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur-R R, Li CH, Becker L, Bellorae N, Zhao XM, Li XF, Wang M, Fang C, Xie B, Zhou ZC, Huang H, Chen SL, Venkatesh B, Shi Q (2018) Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *P Natl Acad Sci* , **115** , 66249-6254.

Jara Z (1988) Some aspects of excretion and osmoregulation of fishes.*Przeglad Zoologiczny* , **32** .

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* , **110** , 462-467.

Kanehisa M, Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* , **28** , 27-30.

Kumar S, Stecher G, Suleski M, Hedges S (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* ,**34** , 1812-1819.

Lehtonen H (1998) Winter biology of burbot (Lota lota L.). *Memo Soc Fauna Flora Fennica* , **74** , 45-52.

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

Liu BH, Shi YJ, Yuan JY, Hu XS, Zhang H, Li N, Chen YX, Mu DS, Fan W (2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *ArXiv preprint arXiv* .

Liu, Q, (2020) Chromosome-level genome assembly of burbot (Lota lota) provides insights into the evolutionary adaptations in freshwater. figshare. Dataset. https://doi.org/10.6084/m9.figshare.12927203.v1

Lobo I (2008) Basic Local Alignment Search Tool (BLAST). *J Mol Biol* , **215** , 403-410.

Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* , **25** , 955-964.

Maitland PS, Lyle AA (1990) Practical conservation of British fishes: current action on six declining species. *J Fish Biol* , **37(Suppl A)** : 319-334.

Mark Chaisson, Glenn Tesler (2012) Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics* , **13** , 238.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNAseq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* , **11** , 1650.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* , **159** , 1665-1680.

Sanderson MJ (2003) r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock.*Bioinformatics* , **19** , 301-302.

Sanetra M, Meyer A (2005) Microsatellites from the burbot (Lota lota), a freshwater gadoid fish (Teleostei). *Mol Ecol Resour* , **5** , 390-392.

Schaefer FJ, Hermelink B, Husmann P, Meeus W, Adriaen J, Wuertz S (2016) Induction of gonadal maturation at different temperatures in burbot Lota lota. *J Fish Biol* , **89** , 2268-2281.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , **31** , 3210-3212.

Stamatakis A (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* ,**30** , 1312-1313.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* , **34** , W435-439.

Stapanian MA, Witzel LD, Cook A (2010) Recruitment of burbot (Lota lota L.) in Lake Erie: An empirical modelling approach. *Ecol Freshw Fish* , **19** , 326-337.

Stapanian MA, Paragamian VL, Madenjian CP, Jackson JR, Lappalainen J, Evenson MJ, Neufeld MD (2010) Worldwide status of burbot and conservation measures. *Fish Fish* , **11** , 34-56.

Star B, Nederbragt A, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzen A, Winer R, Knight J, Vogel J, Aken B, Andersen O, Lagesen K, Tooming-Klunderud A, Edvardsen R, Tina K, Espelund M, Nepal C, Previti C, Karlsen B, Moum T, Skage M, Berg P, Gjoen T, Kuhl H, Thorsen J, Malde K, Reinhardt R, Du L, Johansen S, Searle S, Lien S, Nilsen F, Jonassen I, Omholt S, Stenseth N, Jakobsen K (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature* , **477** , 207-210.

Tammi J, Lappalainen A, Mannio J, Rask M, Vuorenmaa J (1999) Effects of eutrophication on fish and fisheries in Finnish lakes: a survey based on random sampling. *Fisheries Manag Ecol* , **6** , 173-186.

Taşbozan O, Gökçe MA (2017) Fatty Acids in Fish.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf AI, Yin JJ, Natale DA (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* , **4** , 41.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *Plos One* , **9** , e112963.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood.*Mol Biol Evol* , **24** , 1586-1591.

Yao TT, Song L, Jin JJ, Cai Y, Takahashi H, Swanson SK, Washburn MP, Florens L, Conaway RC, Cohen RE, Conaway JW (2008) Distinct modes of regulation of the Uch37 deubiquitinating enzyme in the proteasome and in the Ino80 chromatin-remodeling complex. *Mol cell* , **31** , 909-917.

Zhao R, Diop-Bove N, Visentin M, Goldman ID (2011) Mechanisms of membrane transport of folates into cells and across epithelia.*Annu Rev Nutr* , **31** ,177.

Zhao X, Hao W (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* , **35** , W265-W268.

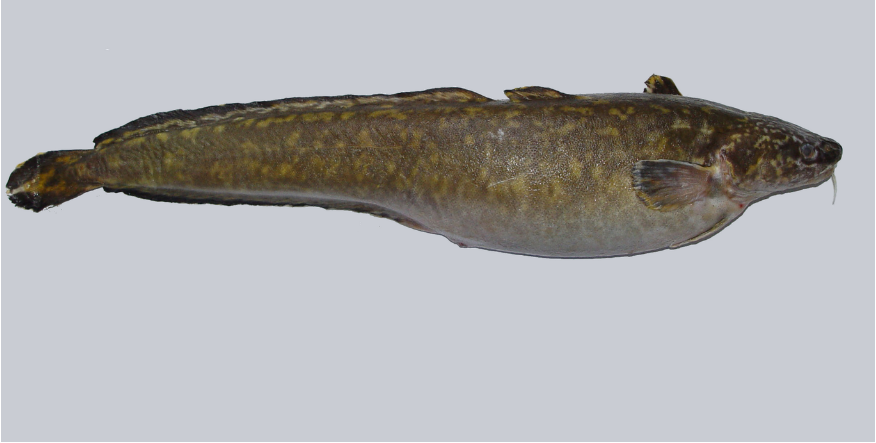
Figure legend

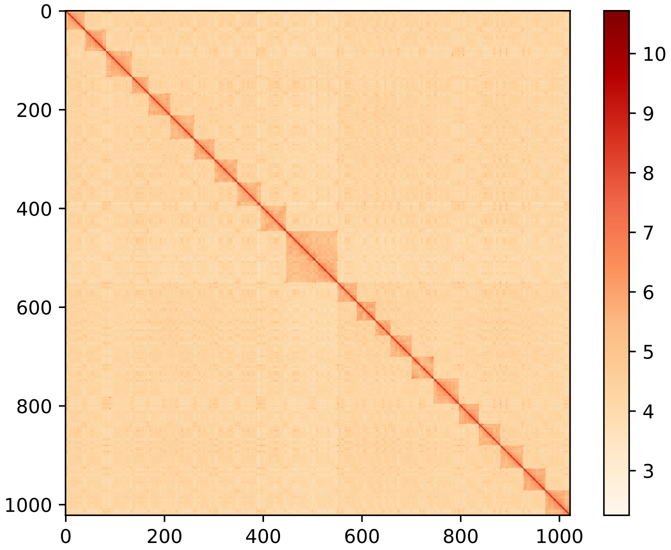
Figure 1 The burbot (*Lota lota*)

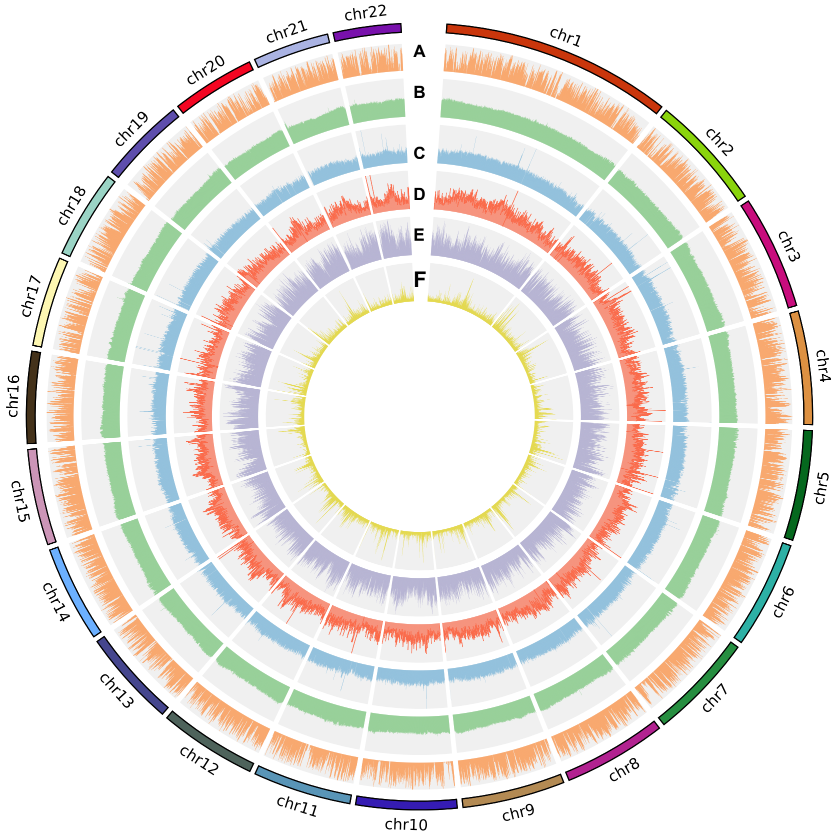
Figure 2 Genome-wide Hi-C heatmap of the burbot.

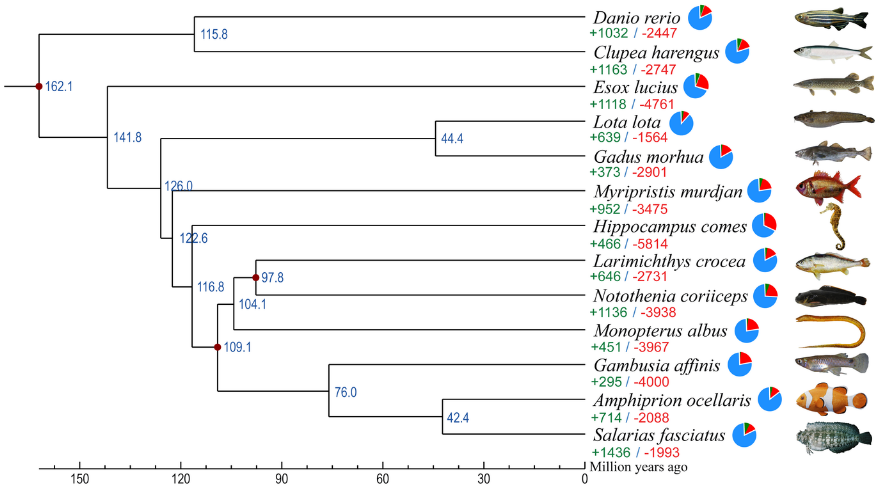
Figure 3 Genome characteristics of burbot. From outer circle to inner circle: gene distribution, GC content of the genome, short read depth, long read depth, DNA TE, and long tandem repeats (LTR).

Figure 4 Phylogenetic analysis and divergence time tree of the burbot with other teleost species.









Hosted file

`tables.pdf` available at https://authorea.com/users/365476/articles/485566-chromosome-level-genome-assembly-of-burbot-lota-lota-provides-insights-into-the-evolutionary-adaptations-in-freshwater