Article

Large-scale phylogenomic analysis provides new insights into the phylogeny of the order Gadiformes and evolution of freshwater gadiform species burbot (*Lota lota*)

Zhiqiang Han [1], Chenyan Shou [1], Manhong Liu [2] and Tianxiang Gao [1,*]

[1] Fishery College, Zhejiang Ocean University, Zhoushan, Zhejiang 316002, China
[2] College of Wildlife &Natural Protected Area, Northeast Forestry University, Harbin, 150040, China
* Correspondence: gaotianxiang0611@163.com

Abstract: Our understanding of phylogenetic relationships among Gadiformes fish is obtained through the analysis of a small number of genes, but uncertainty remains around critical nodes. A series of phylogenetic controversial exists at the suborder, family, subfamily, and species levels. A total of 1105 orthologous exon sequences and translated amino acid sequences from 36 genomes and 12 transcriptomes covering 33 species were applied to investigate the phylogenetic relationships within Gadiformes and address these problems. Phylogenetic trees reconstructed with the amino acid data set using different tree-building methods (RAxML and MrBayes) showed consistent topology. The monophyly of Gadifromes was confirmed in our study. However, the three suborders Muraenolepidoidei, Macrouroidei, and Gadoidei were not well recovered by our phylogenomic study, rejecting the validity of suborder Muraenolepidoidei. Four major lineages were revealed in this study. The family Bregmacerotidae forming clade I was the basal lineage of Gadiformes. The family Merluciidae formed clade II. Clade III contained families Melanonidae, Muraenolepididae, Macrouridae (with subfamilies Trachyrincinae, Macrourinae, and Bathygadinae), and Moridae. Clade IV contained at least three families of suborder Gadoidei, i.e., Gadidae, Phycidae, and Ranicipitidae. The subspecies of *Lota lota* from Amur River were confirmed, indicating that exon markers were a valid high-resolution method for delimiting subspecies or distinct lineages within species level. The PSMC analysis of different populations of *L. lota* suggests a continuous decline since 2 Myr.

Keywords: phylogenomic analyses, exon sequence, Gadiformes, *Lota lota*

1. Introduction

The order Gadiformes includes some of the most important commercial fish in the world, such as cod, hake, and haddock, accounting for approximately a quarter of the world's total marine fish catch [1]. Currently, Gadiformes contains 555 described species in 75 genera and 9 families [1]. Given the commercial and ecological importance of Gadiformes, a longstanding interest in their phylogenetic relationships exists. Beginning with the established by Müller [2], different hypotheses on the relationships of Gadiformes relying on morphological and molecular data have been put forward over the decades. The number of suborders, families, and subfamilies in this order has changed remarkably. The classification of suborders and families in Gadiformes is controversial. Currently, different authors recognized 11–14 families and 2–4 suborders [1, 3]. Recently, a consensus of these classifications is the recognition of three suborders, i.e., Muraenolepidoidei, Gadoidei, and Macrouroidei [4]. However, gadiforms especially the Gadoidei are probably one of the best studied groups of all marine fish worldwide because of their importance to fisheries, but their phylogenetic relationship remains vague [4]. Three authors have raised quite different phylogenetic hypotheses for Gadiformes on the basis of the usage of different morphological characters, such as characters of muscles

and joints in Howes [5], adult morphology in Markle [6], and otolith morphology in Nolf and Steurbaut (1989). Endo [3] has published a comprehensive classification of this order, trying to obtain a consensus on the systematics of Gadiformes (Figure 1a). However, several questions regarding the internal relationships of the major Gadiformes taxa remain due to the homoplasy among morphological characters (Figure 1).

Since the late 1990s, numerous comprehensive analyses within Gadiformes have been published on the basis of molecular data sets. Teletchea et al. [7] have published the phylogeny of the family Gadidae on the basis of morphological characters and two mitochondrial genes and proposed a new provisional classification of gadoids. Roa-Varón and Ortí [4] have conducted a comprehensive molecular phylogeny for Gadiformes on the basis of 2740 bp of DNA sequence from 12S, 16S, and RAG. Results confirm the monophyly of Gadiformes and divide Gadiformes into 12 families in three suborders, i.e., Muraenolepidoidei, Macrouroidei, and Gadoidei, but the placement of Muraenolepididae as sister group to all other Gadiformes is weakly supported. Betancur-R et al., [8] have constructed a new version of the phylogenetic classification of bony fishes covering nearly 2000 species by the mtDNA genome and 20 nuclear genes. Betancur-R et al., [8] show that the family Merlucciidae rather than the family Muraenolepididae is the basal branch in the Gadiformes phylogeny. The placement of Muraenolepididae as sister group to all other Gadiformes is not supported, thereby rejecting the validity of the suborder Muraenolepidoidei. Additionally, the family Lotidae is no longer recognized because it is not monophyletic, and three genera (i.e., *Brosme*, *Lota*, and *Molva*) formerly in Lotidae are now included in Gadidae. However, the families Bregmacerotidae, Euclichthyidae, Melanonidae, Ranicipitidae, Trachyrincidae are not examined by Betancur-R et al., [8]. The topologies of molecular trees and morphological analyses concerning the phylogenetic relationships of Gadiformes are different and contradictory (Figure 1).

We have summarized the incongruence among different studies. The following questions need to be solved. (1) Which family is the basal branch in the Gadiformes phylogeny? Is Muraenolepidae or Merlucciidae the earliest branching lineage among Gadiformes. (2) How many suborders are included in the Gadiformes? (3) How many subspecies exist in Lota lota? Several controversies on the status of subspecies in *L. lota* exist. *L. lota*, the only freshwater member of the cod family, has a circumpolar distribution [4]. Hubbs and Schultz [10] have first raised the idea of three subspecies in *L. lota*, namely, *L. lota leptura*, *L. lota lota*, and *L. lota maculosa*. Pivnicka [11] has performed additional morphological analyses on European burbot populations and concluded two distinct forms of burbot, i.e., *L. lota lota* and *L. lota maculosa*. A molecular survey of cytochrome b polymorphism conducted on 41 populations tried to solve the confusion. Results of cytochrome b are congruent with historically described subspecies, *L. lota lota* and *L. lota maculosa* [12]. However, Fang et al., [13] have reported a new subspecies of *L. lota* from Amur River, China, indicating at least three subspecies of *L. lota*. Fossil evidence suggests that the genus already inhabits European rivers in the early Pliocene [14], indicating that this genus is at least 5 Myr old. However, the molecular survey of cytochrome b reveals that the present-day genetic variation is merely 1 Myr old [12]. The time frame of this unique freshwater gadoid species should be revealed.

Given that mitochondrial genomes represent a single locus and that mitochondrial-based trees are occasionally inaccurate due to introgression, saturation, or selection [15], phylogenetic analyses based on multiple nuclear loci are desirable for elucidating evolutionary history. One promise of the new era of phylogenomics is that increasing number of genes, including whole genomes, can satisfactorily resolve conflicts among few gene studies and improve the robustness and reliability of molecular phylogeny. The Next-generation sequencing drastically reduces the cost of sequencing a genome, and reconstructing phylogenetic relationships by using whole genomes has become feasible [16]. In the last decade, over a thousand genomes have been sequenced, enabling systematics to analyze hundreds to thousands of loci for phylogenetic analysis [17]. Thus, the phylogenomic analysis has become mainstream to resolve phylogenetic relationships

[18]. Recent studies reveal that genome-scale information can unambiguously resolve incongruences in molecular phylogenies. However, the massive data sets generated by high-throughput sequencing especially whole genome data sets require complex analytical methods to confirm the most informative orthologous loci and appropriate tree reasoning methods. Many types of coding and noncoding genomic loci, such as ultraconserved elements (UCEs), introns, or conserved nonexonic elements (CNEEs), are available for reconstructing phylogeny [19], but exons are usually selected to reconstruct fish phylogeny at large scales because of easy alignment and conservation among species [20]. Malmstrøm et al., [21] have performed the phylogenomic analyses of teleost fishes (including 27 Gadiformes fish species) by using a data set of 567 exon orthologs from 111 genes. However, genetic markers are normally selected using similarity without applying stringent analyses to determine the orthology beyond similarity-based criteria. The gene genealogy interrogation can be used to mitigate the effects of gene tree estimation error and filtered paralogs [18]. A new set of exons (including 1105 exons) successfully used for fish phylogenetics is obtained through whole-genome sequencing and gene genealogy interrogation to determine orthology, thereby avoiding paralogy. Hughes et al., [22] have applied this genome-scale data set, conducted the large-scale phylogenomic analysis of 303 fish species, and settled some longstanding uncertainties (such as the branching order at the base of teleosts and among early euteleosts and the sister lineage to the acanthomorph and percomorph radiations). The above studies have confirmed the validity of exon markers for phylogenetic analysis among species, but the validity of this method for delimiting subspecies or distinct lineages within species is unknown.

In this study, we have conducted a comprehensive molecular phylogenetic study of the order Gadiformes, including available genomic and transcriptomic data, by using genome-scale orthologs conserved exon markers. *L. lota* samples from the Irtysh River, China; Amuer River, Chana; and Great Lakes, Canada are included in this study. First, we use phylogenomic analysis to resolve the longstanding controversy regarding the relationships of Gadiformes fish. Second, the status of new subspecies in *L. lota* from Amuer River is checked using the genome-wide locus.
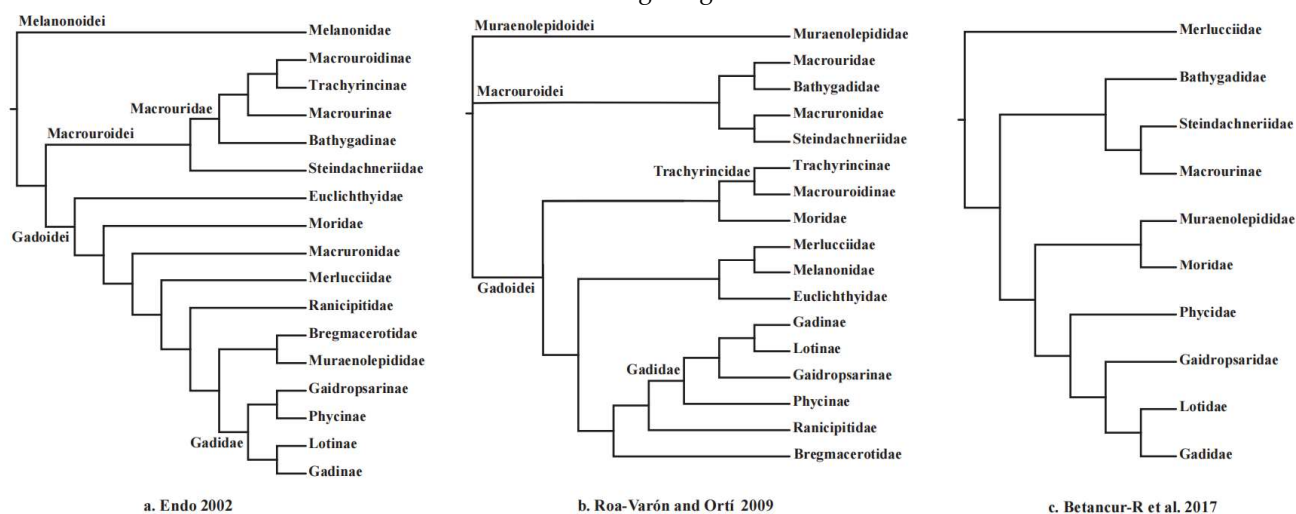

Figure 1. Main hypotheses for Gadiformes proposed by (a) Endo (2002), (b) Roa-Varón and Ortí (2009), and (c) Betancur-R et al. (2017).

2. Results

A total of 36 genomes and 12 transcriptomes, including two newly sequenced genomes from *L. lota* that represented Irtysh River (China) and Great Slave Lake (Canada) populations (Table 1), were used in this study. For each new sample, a single paired-end library was sequenced to 49.5–56.5× coverage on the Illumina NovaSeq 6000 platform. Genome sequences were assembled using the SOAPdenovo2, resulting in N50 config

sizes between 1.8 and 3.2 kb (Table 2). On average, we recovered 48% of the conserved genes of the Actinopterygii data set (3640 genes) included in the BUSCO analysis. Collectively, these results indicated a middle degree of gene space completeness for gene detection in our partial draft genome assemblies. The Trinity software was applied to the de novo assembly of transcriptome data, and 106 084–353 261 unigenes were generated with N50 varying from 319 to 2285. The transcriptome assembly information is shown in Table 1.

We used a set of 1105 single-copy conserved exons for phylogenomic analyses to establish gadiform relationships firmly. Exon alignments for these eight model species were used to obtain HMMs and search for each locus in 36 genomes and 12 transcriptomes, covering 33 species. Significant hits were added to each original alignment to create an exon sequence database for 48 individuals. The concatenated alignment of 1105 exons produced a data matrix with 563 139 bp (187 713 amino acids) for 33 species.

We compared the levels of among- and within-species variations for exon and amino acid data sets (Tables S1 and S2). Tables 3 and 4 showed the genetic distances for selected species (i.e., *Gadus morhua*, *G. chalcogrammus*, and *L. lota*), which were calculated as absolute pairwise DNA and amino acid differences, and related genetic distances. For the pairwise comparisons within species, the ranges of nucleotide and amino acid differences within Atlantic cod (*G. morhua*) were 2226–2779 (Table 3) and 716–920 (Table 4), respectively, with averages of 2530.25 and 804.5, respectively. The range of nucleotide differences within *L. lota* showed high values, which varied from 438 to 2841. Nucleotide differences within the Amur River population of *L. lota* varied from 438 to 1820 (average = 923). The smallest intergeneric difference was observed between *Gadus chalcogrammus* and *G. morhua* (average nucleotide differences, 2530.26; amino acid differences, 804.5), and this finding was similar to the within-species difference between Amur River and EB lineages of *L. lota* (nucleotide substitutions, 1319–2841; amino acid differences, 413–1060).

The phylogenetic analyses reconstructed from amino acid sequences data set by using two tree-building methods, including RAxML and MrBayes, all recovered the same topology (Figures 2 and 3). Our results suggested that Gadiformes could be split into four major lineages that generally corresponded to the families Bregmacerotidae and Merlucciidae and suborders Macrouroidei and Gadoidei. A sister-group relationship between Bregmacerotidae and all other gadiform families was reported. The hypothesis that proposed Muraenolepididae as the independent suborder Muraenolepidoidei and first diverging family was not supported by our analyses. The family Muraenolepididae was clustered into clade III and had a sister-group relationship with Trachyrincinae (*Trachyrincus scabrous* and *Trachyrincus murrayi*). Phylogenetic analyses placed Merlucciidae as the second branching lineage of Gadiformes (clade II). Clade III contained the families Muraenolepididae, Melanonidae, Moridae, and Macrouridae and subfamily Trachyrincinae, including 10 species. This clade covered the previous suborders Muraenolepidoidei and Macrouroidei and two families of suborder Gadoidei (i.e., Trachyrincinae and Moridae). Clade IV included the other families within suborder Gadoidei. The families Gadidae and Phycidae were recovered as paraphyletic groups. Three species (i.e., *L. lota*, *M. molva*, and *Brosme brosme*) in the subfamily Lotinae were analyzed. However, subfamily Lotinae was not recovered as a monophyletic group. *M. molva* and *B. brosme* were more closely related to Gadinae than to Lota.

Assuming a generation length of four years and mutation rate of 0.6×10^{-8}, the demographic histories of *L. lota* encompassed the period from 4 million years ago (Mya) to 10 thousand years ago (Kya) (Figure 4). The effective size (Ne) of all populations peaked at 2 Mya and subsequently declined. The Great Slave Lake population showed two phases of decline at 2 Mya to 0.4 Mya and 30 Kya to 12 Kya, a stable phase from 0.4 Mya to 30 Kya, and pronounced expansion since 12 Kya. High similarities of demographic histories of *L. lota* from Germany, Irtysh River, and Amur River were observed. The first decline of these three populations was from 2 Mya to 0.4 Mya. The second decline of three populations varied remarkably (Germany and Irtysh River: 0.2 Mya to 13

Kya, Amur River: 60 Kya to 10 Kya). The Ne values of Germany and Irtysh River showed slight expansion since 13 Kya.
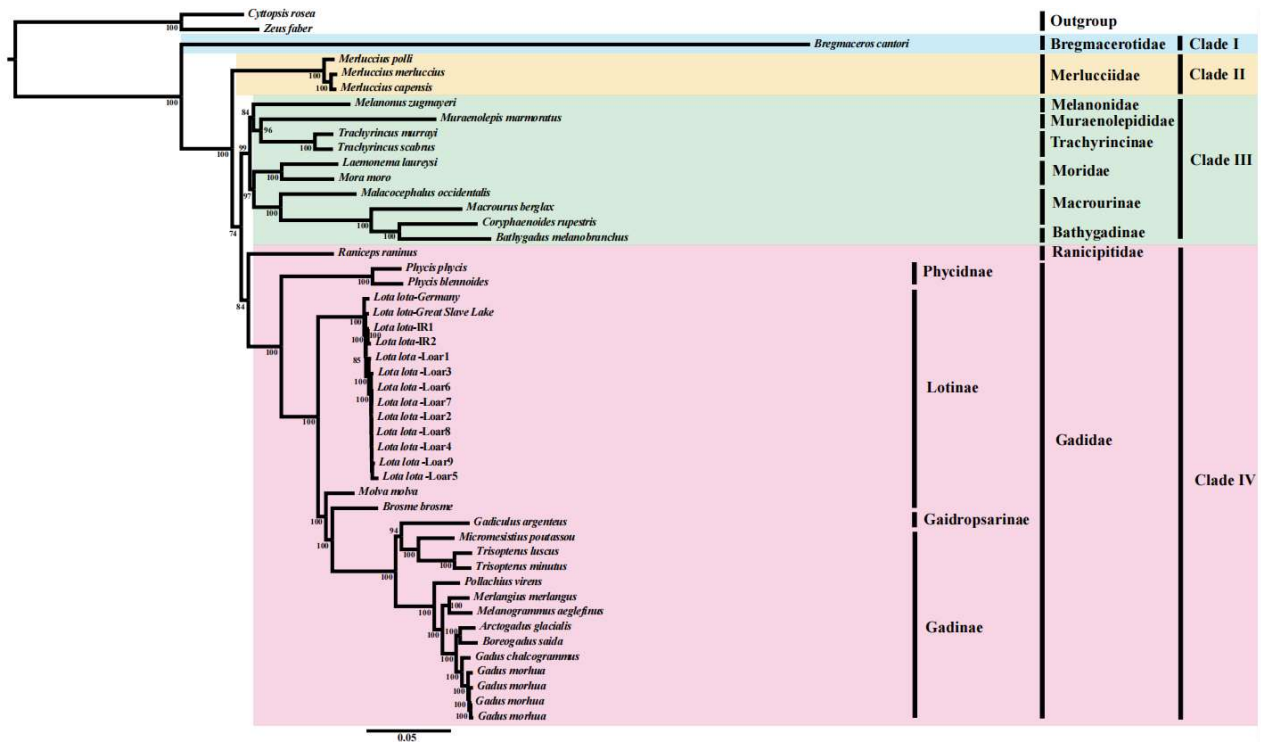

Figure 2. Maximum likelihood tree for 33 taxa based on concatenated amino acid data. Numbers by the nodes indicate ML bootstrap support values (only bootstrap values > 60% are shown).

Figure 3. Bayesian tree for 33 taxa based on concatenated amino acid data. Numbers by the nodes indicate ML bootstrap support values (only bootstrap values > 60% are shown).
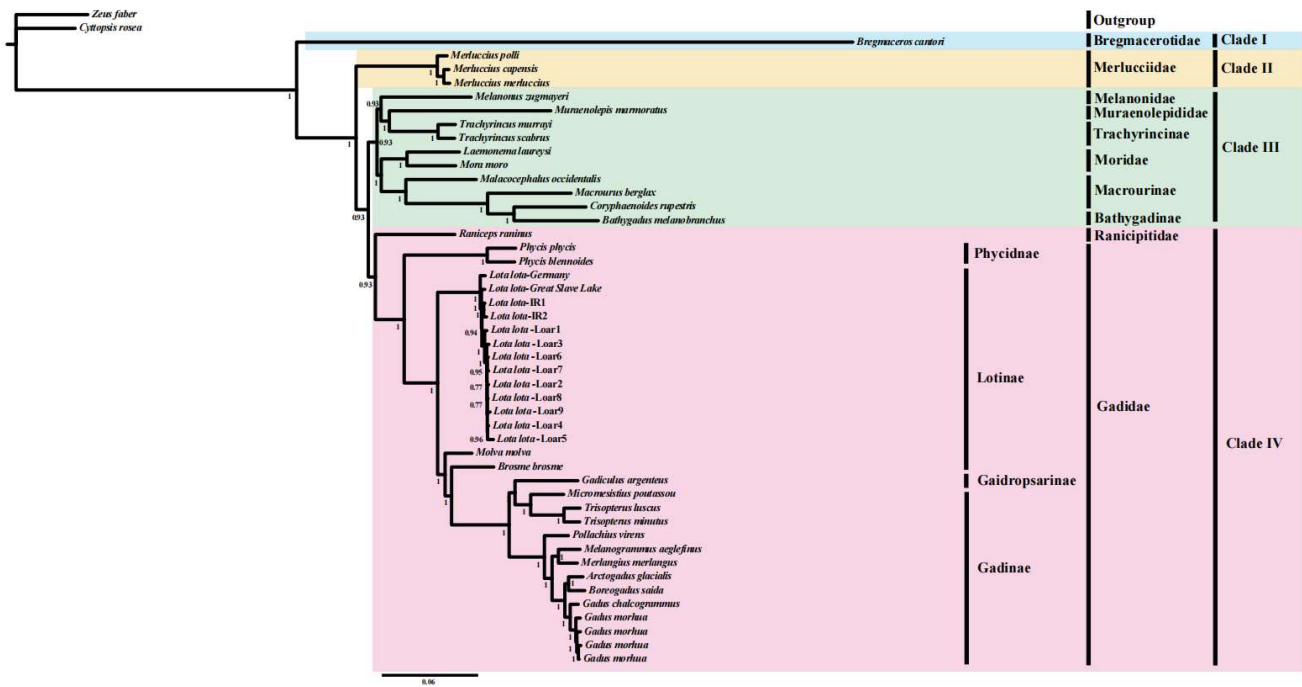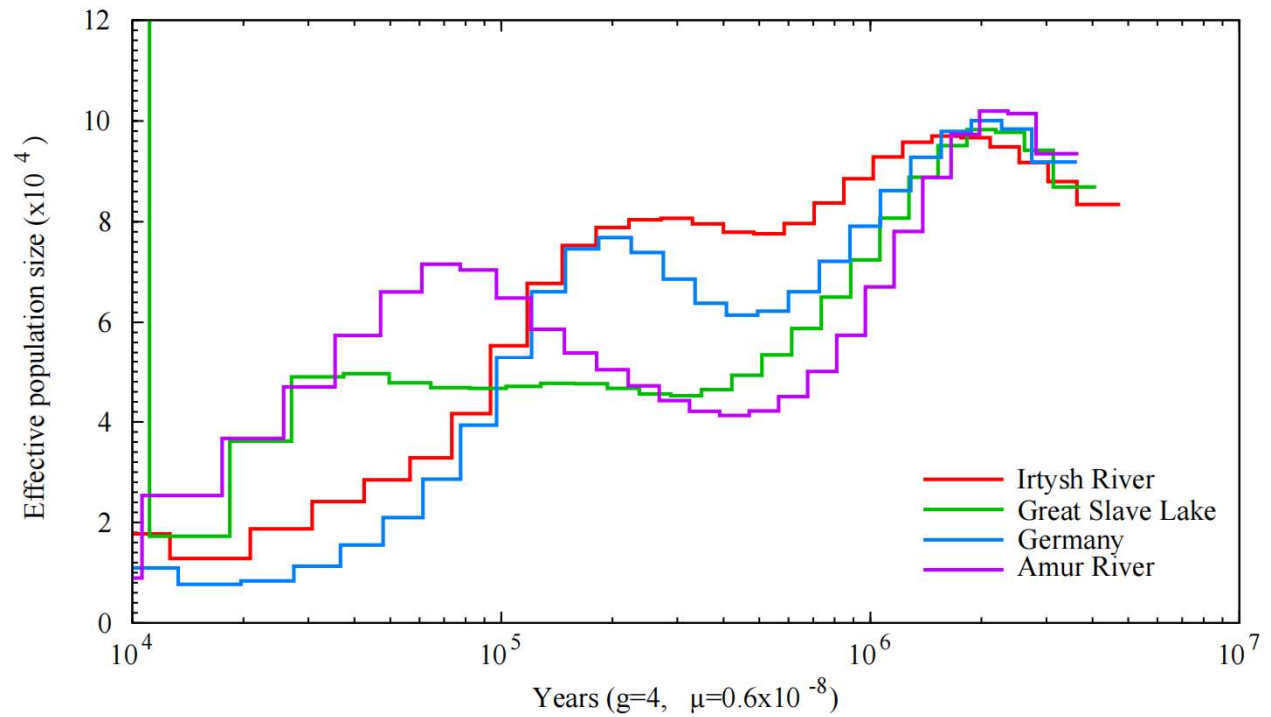
Figure 4. Demographic histories of burbot from different locations. The PSMC estimates of demographic changes in effective population size (Ne) over time are inferred from the genome sequences of burbot.

Table 1. Genomic and transcriptomic data information for the species used in this study. The taxonomy followed the classification of Gadiformes suggested by Roa-Varón and Ortí (2009)

Species	Data type	Total sequence length	Contig N50	Source
1 Muraenolepididae				
Muraenolepis marmoratus	Genome	416,390,766	3,126	GCA_900302325.1
2 Macrouridae				
Macrourinae				
Coryphaenoides rupestris	Genome	829,208,733	20,848	GCA_002895965.1
Malacocephalus occidentalis	Genome	431,202,967	4,956	GCA_900302375.1
Macrourus berglax	Genome	350,339,566	3,697	GCA_900312585.1
Bathygadinae				
Bathygadus melanobranchus	Genome	399,875,629	3,353	GCA_900302365.1
Trachyrincinae	Genome			
Trachyrincus scabrus	Genome	369,861,760	3,900	GCA_900303215.1
Trachyrincus murrayi	Genome	452,416,606	6,231	GCA_900323305.1
3 Moridae				
Laemonema laureysi	Genome	306,494,646	3,431	GCA_900303225.1
Mora moro	Genome	344,961,111	3,267	GCA_900303205.1

4 Merlucciidae

Merluccius capensis	Genome	414,317,329	3,792	GCA_900312945.1
Merluccius merluccius	Genome	401,034,705	3,670	GCA_900312545.1
Merluccius polli	Genome	401,149,128	3,471	GCA_900312625.1

5 Melanonidae

Melanonus zugmayeri	Genome	432,902,915	4,562	GCA_900302345.1

6 Gadidae

Gadinae

*Gadus morhua*1	Genome	669,966,409	1,015,663	GCA_902167405.1,GM1
*Gadus morhua*2	Genome	684,284,792	10,559,872	GCA_010882105.1,GM2
*Gadus morhua*3	Genome	495,065,230	5,765	GCA_900302565.1,GM3
*Gadus morhua*4	Genome	824,311,139	2,311	GCA_000231765.1,GM4
Melanogrammus aeglefinus	Genome	652,790,733	77,605	GCA_900291075.1
Gadus chalcogrammus	Genome	448,868,398	3,603	GCA_900302575.1,GC
Merlangius merlangus	Genome	423,942,190	3,538	GCA_900323355.1
Pollachius virens	Genome	394,927,939	3,457	GCA_900312635.1
Arctogadus glacialis	Genome	428,791,846	3,282	GCA_900303235.1
Trisopterus minutus	Genome	334,717,091	3,248	GCA_900302415.1
Boreogadus saida	Genome	412,070,465	3,221	GCA_900302515.1
Micromesistius poutassou	transcriptome	141,831,209	929	SRR9822077
Trisopterus luscus	transcriptome	123,118,811	925	SRR9822078

Gaidropsarinae

Gadiculus argenteus	Genome	396,767,394	3,379	GCA_900302595.1

Lotinae

Molva molva	Genome	437,480,619	4,136	GCA_900323295.1
Brosme brosme	Genome	412,731,310	3,682	GCA_900302425.1

Lota lota

Lota lota lota (EB lineage)

Lota lota-Germany	Genome	397,499,185	3,803	GCA_900302385.1,collected from Germany, LL-G
Lota lota-IR1	Genome	505,651,266	3,231	This study, collected from Irtysh River, China, LL-I1
Lota lota-IR2	transcriptome	75,050,706	2,285	Yang et al., 2020,collected from Irtysh River, China,LL-I2

Lota lota maculosa (NA lineage)

Lota lota-Great Slave Lake	Genome	583,659,647	1,882	This study,collected from Great Slave Lake, Canda, LL-GSL

Amur River lineage

Lota lota -Loar1	Genome	575,834,103	2,146,850	Han et al.,2021,collected from,Amur River, China, LL-L1
Lota lota -Loar2	transcriptome	99,174,918	1,681	SRR13435971,collected from,Amur River, China, LL-L2
Lota lota -Loar3	transcriptome	105,108,511	1,499	SRR13435972,collected from,Amur River, China, LL-L3
Lota lota -Loar4	transcriptome	105,556,503	1,572	SRR13435973,collected from,Amur River, China, LL-L4
Lota lota -Loar5	transcriptome	100,225,292	1,634	SRR13435974,collected from,Amur River, China. LL-L5
Lota lota -Loar6	transcriptome	113,753,425	1,520	SRR13435975,collected from,Amur River, China. LL-L6
Lota lota -Loar7	transcriptome	106,928,281	1,569	SRR13435976,collected from,Amur River, China. LL-L7
Lota lota -Loar8	transcriptome	109,461,422	1,522	SRR13435977,collected from,Amur River, China. LL-L8
Lota lota -Loar9	transcriptome	91,753,235	1,769	SRR13435978,collected from,Amur River, China. LL-L9

Phycidnae

　Phycis blennoides 　　Genome　　416,766,999　　4,532　　GCA_900302315.1

　Phycis phycis 　　Genome　　346,335,180　　3,458　　GCA_900302335.1

7 Ranicipitidae

　Raniceps raninus 　　transcriptome　　113,188,970　　319　　SRR12432352

8 Bregmacerotidae

　Bregmaceros cantori 　　Genome　　1,144,104,325　　4,452　　GCA_900302395.1

Outgroup

Zeiformes

　Cyttopsis rosea 　　Genome　　545,311,884　　4,843　　ASM90030235v1

　Zeus faber 　　Genome　　609,364,180　　4,642　　ASM90032333v1

Table 2. Statistics for the sequencing data of burbot draft genomes

	Library type	Read number	Total length (Gb)	Coverage	Q20(%)	Complete*	Duplicated*	Fragmented*	Missing*
Lota lota (Irtysh River1)	300-bp pair-end	107,926,805	16.05	29.18	94.32	58.6	0.6	12	28.8
Lota lota (Great Slave Lake)	300-bp pair-end	191,982,138	28.35	51.54	96.51	36.7	0.3	13.3	49.7

[2]out of 3640 highly conserved acantopterygii genes based on BUSCO.

Table 3. Pairwise numbers of nucleotide differences (below) and genetic distance (above) based on the exon sequences of three species.

	GM1	GM2	GM3	GM4	GC	LL-G	LL-I1	LL-I2	LL-GSL	LL-L1	LL-L2	LL-L3	LL-L4	LL-L5	LL-L6	LL-L7	LL-L8	LL-L9
GM1		0.0048	0.0026	0.0028	0.0086	0.0989	0.0983	0.0974	0.0974	0.0986	0.0990	0.0991	0.0981	0.1009	0.0981	0.0985	0.0984	0.0995
GM2	2111		0.0039	0.0036	0.0085	0.1000	0.0995	0.0976	0.0984	0.0999	0.0990	0.0995	0.0987	0.1011	0.0987	0.0989	0.0989	0.0998
GM3	947	1450		0.0017	0.0071	0.0966	0.0961	0.0952	0.0955	0.0967	0.0959	0.0957	0.0953	0.0978	0.0951	0.0958	0.0957	0.0969
GM4	1081	1478	642		0.0074	0.0958	0.0955	0.0946	0.0951	0.0960	0.0955	0.0952	0.0945	0.0972	0.0949	0.0954	0.0951	0.0959
GC	2779	2759	2226	2357		0.0942	0.0935	0.0930	0.0934	0.0938	0.0943	0.0945	0.0932	0.0951	0.0933	0.0945	0.0937	0.0942
LL-G	32929	33905	27825	29762	24917		0.0041	0.0051	0.0043	0.0047	0.0058	0.0057	0.0053	0.0079	0.0054	0.0058	0.0056	0.0065
LL-I1	35555	37260	29588	32113	25463	1664		0.0022	0.0032	0.0035	0.0042	0.0046	0.0040	0.0063	0.0040	0.0044	0.0041	0.0049
LL-I2	27986	29061	23589	25692	20628	1641	779		0.0039	0.0050	0.0040	0.0044	0.0038	0.0051	0.0037	0.0042	0.0037	0.0047
LL-GSL	33908	35476	28666	31148	24900	1700	1373	1341		0.0037	0.0044	0.0047	0.0042	0.0064	0.0040	0.0046	0.0042	0.0050
LL-L1	36004	37073	29380	31843	25739	1940	1554	1763	1588		0.0028	0.0032	0.0025	0.0046	0.0023	0.0027	0.0025	0.0034
LL-L2	31442	32372	25657	28073	22453	2064	1668	1378	1669	1114		0.0021	0.0015	0.0038	0.0014	0.0015	0.0014	0.0020
LL-L3	31734	32898	26113	28413	22841	2033	1827	1541	1808	1270	842		0.0019	0.0040	0.0018	0.0018	0.0017	0.0027
LL-L4	32008	33208	26240	28442	23042	1949	1621	1326	1629	1010	576	754		0.0033	0.0013	0.0013	0.0012	0.0018
LL-L5	32316	33492	26472	28920	23009	2841	2512	1791	2434	1820	1482	1566	1312		0.0034	0.0032	0.0034	0.0036
LL-L6	32782	34098	26899	29365	23504	2018	1666	1329	1613	961	568	722	526	1377		0.0012	0.0011	0.0020
LL-L7	32337	33380	26567	28912	23450	2121	1770	1462	1789	1109	579	719	530	1294	479		0.0011	0.0019

LL-L8	32723	33854	26965	29268	23740	2081	1694	1319	1648	1018	548	697	499	1352	458	438		0.0020
LL-L9	31357	32546	25919	27999	22589	2291	1913	1587	1891	1336	773	1040	706	1397	811	760	782	

Table 4. Pairwise number of amino acid differences (below) and genetic distances (above) based on the amino acid sequences of three species.

	GM1	GM2	GM3	GM4	GC	LL-G	LL-I1	LL-I2	LL-GSL	LL-L1	LL-L2	LL-L3	LL-L4	LL-L5	LL-L6	LL-L7	LL-L8	LL-L9
GM1		0.0049	0.0026	0.0028	0.0086	0.0916	0.0907	0.0885	0.0903	0.0918	0.0906	0.0908	0.0905	0.0945	0.0907	0.0904	0.0906	0.0924
GM2	716		0.0038	0.0031	0.0083	0.0928	0.0920	0.0885	0.0913	0.0929	0.0901	0.0910	0.0910	0.0946	0.0911	0.0906	0.0909	0.0927
GM3	305	467		0.0015	0.0065	0.0868	0.0857	0.0837	0.0853	0.0866	0.0846	0.0843	0.0843	0.0885	0.0846	0.0849	0.0853	0.0867
GM4	362	421	179		0.0067	0.0858	0.0851	0.0831	0.0851	0.0859	0.0842	0.0837	0.0834	0.0877	0.0845	0.0843	0.0843	0.0854
GC	920	897	685	716		0.0836	0.0818	0.0799	0.0823	0.0831	0.0823	0.0822	0.0817	0.0847	0.0818	0.0830	0.0829	0.0829
LL-G	10684	11015	8781	9367	7772		0.0042	0.0053	0.0046	0.0049	0.0060	0.0059	0.0056	0.0090	0.0058	0.0063	0.0058	0.0070
LL-I1	11507	12073	9267	10044	7836	569		0.0020	0.0032	0.0034	0.0042	0.0047	0.0040	0.0071	0.0040	0.0045	0.0040	0.0051
LL-I2	8916	9243	7283	7921	6233	564	239		0.0040	0.0051	0.0038	0.0043	0.0036	0.0055	0.0036	0.0042	0.0035	0.0047
LL-GSL	11012	11517	8999	9779	7712	595	461	456		0.0038	0.0045	0.0050	0.0044	0.0073	0.0042	0.0049	0.0043	0.0053
LL-L1	11745	12076	9238	9997	8008	670	509	595	542		0.0031	0.0035	0.0027	0.0053	0.0025	0.0029	0.0026	0.0036
LL-L2	10071	10321	7940	8685	6868	707	547	431	564	396		0.0022	0.0013	0.0044	0.0014	0.0014	0.0012	0.0021
LL-L3	10197	10545	8084	8782	6983	701	616	502	632	450	281		0.0020	0.0047	0.0019	0.0019	0.0017	0.0029
LL-L4	10339	10712	8144	8813	7084	677	539	420	567	353	175	264		0.0038	0.0013	0.0012	0.0011	0.0019
LL-L5	10568	10938	8387	9139	7168	1060	933	634	920	689	561	607	503		0.0041	0.0038	0.0040	0.0042
LL-L6	10607	11027	8402	9170	7232	716	555	421	550	338	181	250	176	548		0.0011	0.0010	0.0021
LL-L7	10413	10728	8266	8977	7230	763	610	486	637	388	178	246	166	500	147		0.0010	0.0021
LL-L8	10563	10905	8436	9108	7366	717	551	413	564	350	163	227	151	526	137	136		0.0019

LL-L9	10192	10567	8135	8743	6970	817	659	531	669	467	262	375	247	542	281	276	255

3. Discussion

3.1. Phylogenetic relationships of order Gadiformes

In this study, we have tried to construct the phylogeny of Gadiformes at the genome level to solve major phylogenetic questions among gadiform fish. The monophyly of Gadiformes is confirmed in our study. However, most formerly described suborders in Gadiformes are not supported by our phylogenomic study.

Results from our genome data set and Malmstrøm et al., [21] suggest that Gadiformes can be split into four major lineages that do not correspond to the three previously described suborders, namely, Muraenolepidoidei, Macrouroidei, and Gadoidei. The suborder Muraenolepidoidei is first established by Svetovidov [23] on the basis of the peculiar structure of the pectoral girdle. The suborder Muraenolepidoidei is thought to represent an ancient lineage of Gadiformes [24]. By contrast, other authors have assigned muraenolepids as a family to the suborder Gadoidei [3]. A recent study has found that its placement as basal lineage of Gadiformes is weakly supported [4]. In our result, the placement of family Muraenolepidoidei as sister group to Trachyrincinae is strongly supported. Further studies should clarify if Muraenolepididae represents the most basal branch in the Gadiformes phylogeny.

Clade I only contains the family Bregmacerotidae, which is the basal lineage of Gadiformes in our study. The phylogenetic position of Bregmacerotidae has been included in the gadoids or regarded as the suborder Bregmacerotoidei [25]. Roa-Varón and Ortí (2009) have reported a discrepancy between mtDNA and nuclear DNA for the placement of Bregmacerotidae. Mitochondrial DNA data show the close relationship of Bregmacerotidae with Macrourinae. The combined data set of mtDNA and nuclear genes in Roa-Varón and Ortí (2009) showed the family Bregmacerotidae in a monophyletic clade with Ranicipitidae in the suborder Gadoidei. Our analyses suggest the position of Bregmacerotidae as the sister group of all other gadiforms. This family may represent the most basal branch in the Gadiformes phylogeny. The unexpected depth of this divergence suggests that Bregmacerotidae can be ranked at the suborder level. However, the long branches separating Bregmaceros from the other taxa in all analyses suggest a rapid evolutionary rate of changes that can saturate the phylogenetic signal of genes. This long branch may mislead the interpretation of phylogenetic information. Similar long branches of Bregmacerotidae are also reported in mtDNA genes [4]. Further study should clarify if Bregmacerotidae represents the most basal branch in the Gadiformes phylogeny.

The family Merluciidae forms clade II in our study. Several taxonomic revisions about the placement of Merluciidae have been published [26-27]. The controversy focuses on whether Merlucciidae is a separate family within Gadoidei. Our genomic data presents new evidence for considering Merlucciidae as a new independent suborder within gadiforms, representing the basal branch in the Gadiformes phylogeny. This result is anticipated at least in part by Betancur-R et al., [8], who have proposed that the family Merluciidae is the most basal suborder.

Clade III contains some families from suborders Macrouroidei and Gadoidei. These families include Melanonidae, Muraenolepididae, Macrouridae (with subfamilies Trachyrincinae, Macrourinae, and Bathygadinae), and Moridae. Three of four subfamilies in Macrouridae are included in the present study, and three monophyletic subclades are recovered in clade III. However, the monophyly of Macrouridae is not well supported, thereby rejecting the monophyly of this family reported in Betancur-R et al., [8] on the basis of the mtDNA genome. Only subfamilies Macrourinae and Bathygadinae are sister groups within Macrouridae. The monophyly relationship between Macrourinae and Bathygadinae is also supported by morphological study and molecular markers [4, 24, 28]. Another nominal subfamily Trachyrincinae belonging to Macrouroidei seems to

be distantly related and clusters with Muraenolepididae. The subfamily Trachyrincinae and families Melanonidae and Moridae are placed in the suborder Gadoidei [4]. In agreement with our study, Cohen [9] adds Moridae and Trachyrincidae in the suborder Macrouroidei. The placement of the family Melanonidae is largely incongruent among previous studies. The phylogenetic position of this family is controversial. The family Melanonidae is included in the suborder Gadoidei or forms an independent suborder as the sister group of all Gadiformes [3, 6]. Melanonidae as an independent family within Macrouroidei is strongly supported in our study.

Clade IV contains at least three families of the suborder Gadoidei, i.e., Gadidae, Phycidae, and Ranicipitidae. The suborder Gadoidei is not well recovered by the genomic data set and has been defined by two principal clades in Roa-Varón and Ortí [4]. The first clade, which contains Moridae and two subfamilies of Macrouridae (i.e., Trachyrincinae and Macrouroidinae), is moved to clade III in our study. The second clade, which includes Merlucciidae, Melanonidae, Euclichthyidae (lack of sample in our study), Ranicipitidae, Bregmacerotidae, and Gadidae, forms clade IV of our results. The results obtained with our genomic data set suggest that the suborder Gadoidei contains at least three families, namely, Gadidae, Phycidae, and Ranicipitidae. The family Gadidae contains some commercially important species and has attracted many studies. Previous morphological studies prove the validity of the family Gadidae. Our data support a monophyletic clade for taxa in the family Gadidae with high bootstrap values. However, two subfamilies Lotinae and Gaidropsarinae are not recovered as monophyletic groups. The three species of the subfamily Lotinae (i.e., *L. lota*, *M. molva*, and *B. brosme*) are included in the present study. *M. molva* and *B. brosme* in the present study are more closely related to Gadinae than to *L. lota*. A previous study has shown the nonmonophyletic relationship between *L. lota* and *M. molva* and suggested to clarify the status of Lotinae by including *B. brosme* to cover all species in the subfamily. Notably, the different positions of *L. lota* are revealed and compared with those observed by Roa-Varon and Ortí [4]. Roa-Varon and Ortí [4] have revealed that *L. lota* is more closely related to Gadinae than to *M. molva*. However, our result indicates the early divergence for Lota genus and a relatively distant relationship to Gainae, thereby revising previous finding. Therefore, *M. molva* and *B. brosme* may be moved to the subfamily Gadinae. Additionally, *Gadiculus argenteus*, which represents the subfamily Gaidropsarinae, is the sister group to *Trisopterus* and *Micromesistius* from the subfamily Gadinae. This finding does not recover the subfamily Gaidropsarinae and suggests its inclusion to Gadinae. The status of the subfamily Gaidropsarinae needs revision. More taxa from the subfamily Gaidropsarinae are needed for further study.

3.2. Subspecies status of L. lota

Previous morphological and molecular marker studies on the number of subspecies present in *L. lota* have been controversial [12]. One possible new species or subspecies is confirmed in our genomic study. Hubbs and Schultz [10] have raised the idea of three subspecies on the basis of morphological variation, namely, *L. lota lota*, *L. lota leptura*, and *L. lota maculosa*. The validity of *L. lota lota* and *L. lota maculosa* is confirmed using mtDNA markers [12]. The *L. lota lota* (EB lineage) taxon lives in Eurasia and Canada, and the *L. lota maculosa* (NA lineage) taxon lives in southern Canada and United States. New subspecies of *L. lota* from Amur River, China first described by Fang et al., [13] is confirmed in our genomic data. The average divergence between Amur River and other burbot lineages is within the range of interspecific divergence between *G. chalcogrammus* and *G. morhua*, supporting the status of Amur River as subspecies or even new species. The evolutionary mtDNA lineage of burbot already exists in the Early Pliocene with PSMC results, rendering sufficient time for speciation. Considering the worldwide distribution, more subspecies or species may exist in *L. lota*.

The results of our study provide novel insights into the demographic history of burbot populations. On the basis of the PSMC results, the demographic history of *L. lota* is estimated back up to 4 Mya. This time frame is consistent with that of previous studies.

mtDNA markers and fossil records provide evidence that the genus inhabits freshwater as long as 5 Mya [12, 14]. The PSMC can be used to infer the timing of divergence among distinct populations under a wide range of demographic scenarios. Our PSMC-based estimates of divergence times (about 2 Myr) among the lineages of *L. lota* are similar to previous estimates produced using different molecular markers (about 1.1–1.9 Mya) [12]. A striking difference in demographic histories is observed between the Great Slave Lake population and other populations. The Great Slave Lake population shows gradual and sudden expansion after the last glaciation. This finding corresponds well with the strong signal of ice age in the North America. Other populations show long-term demographic contractions until the end of last glacial. The Irtysh River and Germany populations show the most similar demographic histories, reflecting exposure to similar climatic history.

4. Materials and Methods

4.1. Genomic and transcriptomic data set collection

Genomic and transcriptomic data for 40 species representing 8 gadiform families and 2 outgroup species (i.e., *Cyttopsis rosea* and *Zeus faber*) from the order Zeiformes were retrieved from the NCBI database (Table 1). A new genome assembly of *L. lota* from Amur River, China was included in this study [29].

4.2. New genome sequencing and assembly

Two genome assemblies of *L. lota* were newly sequenced in this study. Samples were collected from Irtysh River, China and Great Slave Lake, Canada, representing the subspecies *L. lota lota* and *L. lota maculosa*, respectively. The subspecies were identified using the mtDNA cytb gene extracted from the draft genome assembly data. Genomic DNA was obtained from the muscle in this study. A single paired-end library with an insert size of ~300 bp was created for each sample by using the Illumina protocol. Two samples were sequenced (2 × 150 bp) to about 50× coverage on the Illumina NovaSeq 6000 platform, and sequences were assembled using the SOAPdenovo2 (https://github.com/aquaskyline/SOAPdenovo2). The draft genome assembly quality in terms of gene space completeness was assessed using the BUSCO 5.0 [30] (Table 2). Raw sequencing data for *L. lota* genome were deposited at the Sequence Read Archive (SRR**).

4.3. Transcriptome de novo assembly

Twelve raw transcriptome data, including eight *L. lota* individuals from the Amur River and one *L. lota* individual from the Irtysh River, China, were collected from the Sequence Read Archive database of NCBI (SRR13435971-SRR13435978) and reference [31], respectively. The raw RNA-seq data were cleaned by removing reads with sequencing adapters, unknown nucleotides (N ratio > 10%), and low-quality reads (quality scores < 20) by using the Trimmomatic 0.36. The software Trinity (Version 2.4.0) [32] software was used for the transcriptome de novo assembly of each individual, and the parameters were as follows: --genome_guided_max_intron 10000.

4.4. Exon captures and phylogenetic inference

A total of 36 Gadiformes genomes and 12 transcriptomes, representing the major families described in Roa-Varón [4], were included for phylogenomic analysis. The single-copy orthology locus was searched in accordance with that published by Hughes et al. [22]. A data set of 1105 single-copy conserved exon markers > 200 bp was obtained from Hughes et al. [22] by comparing eight well-annotated fish genomes. These exon markers passed the paralogy filtering through the gene genealogy interrogation.

The nHMMER program within the HMMER (v3.3.2) [33] was used to search against each study species data set through the hidden Markov model (HMM). High-quality single-copy conserved nuclear coding sequences of each research species were aligned and spliced into single-nucleotide sequences by using the MACSE software (v2.05) [34] on the basis of their amino acid translation. Gene and translated amino acid alignments were concatenated using SCaFoS [35]. Amino acid sequences for phylogenetic analysis can reduce the confounding effect of base-composition heterogeneity among taxa, a biasing factor shown to be extensive among fishes. The maximum likelihood (ML) and

Bayesian inference were used for phylogenetic analyses. The protest 3 [36] was applied to estimate the best-fit model for the amino acid data set. A total of 100 bootstrap replicates were obtained for the optimal PROTGAMMAIJTT substitution model of all concatenated nucleotide and protein sequences in the RAxML v 8.2.12, [37] for the ML tree. The MrBayes analysis was conducted in the MrBayes 3.2.6 [38] with 1 × 105 metropolis-coupled Markov Chain Monte Carlo (MCMCMC) generations. Sampling was done every 100 generations, and the first 25% of the generations were discarded as burn-in.

4.5. History analysis of L. lota

The demographic history for *L. lota* was analyzed using the Pairwise Sequentially Markovian Coalescent (PSMC) model as implemented in the PSMC package v0.6.5 [39]. This software package could infer the population size history from one diploid sequence by using the PSMC model. The whole-genome diploid consensus sequence was generated as follows. The "BWA-MEM" algorithm from the bwa v0.7.17 [40] was used to map the fastq data to the assembled genome of *L. lota*. The SAMtools v0.1.19 [41] was used to generate the diploid consensus with default settings (https://github.com/lh3/psmc). The default settings of the PSMC were adopted. Mutation rates were estimated along the branches of the phylogeny reported in Malmstrom et al., [37], and generation times were set to four years [9].

References

1. Nelson, J. Fishes of the World. 4th Edition. New Jersey, John Wiley & Sons. 2006.
2. Müller, J. Über den Bau und die Grenzen der Ganoiden und über das natürliche System der Fische. *Archiv Für Naturgeschite* **1845**, *11*, 91-141.
3. Endo, H. Phylogeny of the order Gadiformes (Teleostei, Paracanthopterygii). *Memoirs of the Graduate School of Fisheries Sciences, Hokkaido University* **2002**, *49*, 75-149.
4. Roa-Varón, A.; Ortí, G. Phylogenetic relationships among families of Gadiformes (Teleostei, Paracanthopterygii) based on nuclear and mitochondrial data. Molecular Phylogenetics and Evolution **2009**, 52, 688-704. doi: 10.1016/j.ympev.2009.03.020
5. Howes, G.J. Phylogenetic relationships of macrourid and gadoid fishes based on cranial myology and arthrology. In: Cohen DM. Papers on the Systematics of Gadiform Fishes. Natural History Museum of Los Angeles County. 1989.
6. Markle, D.F. Aspects of character homology and phylogeny of the Gadiformes. In: Cohen DM. Papers on the Systematics of Gadiform Fishes. Natural History Museum of Los Angeles County. 1989.
7. Teletchea, F.; Laudet, V.; Hanni, C. Phylogeny of the Gadidae (sensu Svetovidov, 1948) based on their morphology and two mitochondrial genes. *Molecular Phylogenetics and Evolution* **2006**, *38*, 189-199. doi: 10.1016/j.ympev.2005.09.001
8. Betancur-R, R.; Wiley, E.O.; Arratia, G.; Acero, A.; Bailly, N.; Miya, M.; Lecointre, G.; Ortí, G. Phylogenetic classification of bony fishes. *BMC Evolutionary Biology* **2017**, *17*, 162. doi: 10.1186/s12862-017-0958-3

9. Stapanian, M.A.; Paragamian, V.L.; Madenjian, C.P.; Jackson, J.R.; Lappalainen, J.; Evenson, M.J.; et al. Worldwide status of burbot and conservation measures. *Fish and Fisheries* **2010**, *11*, 34-56. doi: 10.1111/j.1467-2979.2009.00340.x

10. Hubbs, C.L.; Schultz, L.P. Contributions to the ichthyology of Alaska with description of two new fishes. *Occasional Papers of the Museum of Zoology* **1941**, *431*, 1-31. doi: http://hdl.handle.net/2027.42/56870

11. Pivnicka, K. Morphological variation in the burbot (Lota lota) and recognition of the subspecies: a review. *Journal of the Fisheries Research Board of Canada* **1970**, *27*,1757-1765. doi: 10.1139/f70-196

12. Houdt, J.K.V.; Hellemans, B.; Volckaert, F.A.M. Phylogenetic relationships among Palearctic and Nearctic burbot (Lota lota): Pleistocene extinctions and recolonization. *Molecular Phylogenetics and Evolution* **2003**, *29*, 599-612. doi: 10.1016/S1055-7903(03)00133-7

13. Fang, H.H.; Zhang, J.L.; Song, N.; Qian, L.; Gao, T.X. Population genetic structure and geographical differentiation of burbot (Lota lota) in China. *Russian Journal of Genetics* **2013**, *49*, 1047-1056. doi: 10.1134/S102279541307003X

14. Pietschmann, V. Lota hulai, eine neue fischart aus dem Wiener Becken. *Palaeontologische Zeitschrift* **1934**, *16*, 48-52. doi: 10.1007/BF03041666

15. Funk, D.J.; Omland, K.E. Species-level paraphyly and polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics* **2003**, *34*, 397-423. doi: 10.1146/annurev.ecolsys.34.011802.132421

16. 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* **2014**, *346*, 1320-1331. doi: 10.1126/science.1253451

17. Faircloth, B.C. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **2012**, *61*, 717-726. doi: 10.1093/sysbio/sys004

18. Betancur-R, R.; Arcila, D.; Vari, R.P.; Hughes, L.C.; Oliveira, C.; Sabaj, M.H.; Ortí, G. Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. *Evolution* **2019**, *73*. doi: 10.1111/evo.13649

19. Edwards, S.V.; Cloutier, A.; Baker, A.J. Conserved nonexonic elements: A novel class of marker for phylogenomics. *Systematic Biology* **2017**, *66*, 1028-1044. doi: 10.1093/sysbio/syx058

20. Jiang, J.M.; Yuan, H.; Zheng, X.; Wang, Q.; Kuang, T.; Li, J.Y.; et al. Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecology and Evolution* **2019**, *9*, 3973-3983. doi: 10.1002/ece3.5026

21. Malmstrøm, M.; Matschiner, M.; Tørresen, O.K.; Star, B.; Snipen, L.G.; Hansen, T.F.; et al. Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics* **2016**, *48*, 1204-1210. doi: 10.1038/ng.3645

22. Hughes, L.C.; Ortí, G.; Huang, Y.; Sun, Y.; Baldwin, C.C.; Thompson, A.W.; Arcila, D.; Betancur-R., R.; Li, C.H.; Becker, L.; Bellora, N.; Zhao, X.M.; Li, X.F.; Wang, M.; Fang, C.; Xie, B.; Zhou, Z.C.; Huang, H.; Chen, S.L.; Venkatesh, B.; Shi, Q. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 6249-6254. doi: 10.1073/pnas.1719358115

23. Svetovidov, A.M. Gadiformes. In Fauna of USSR, 9(4). Zoological Institute of the Academy of Sciences of the USSR (Translated by the Israel Program for Scientific Translations for the Natural Science Foundation and Smithsonian Institution as OTS 63-11071). 1948.

24. Okamura, O. Studies on the macrourid fishes of Japan, morphology, ecology and phylogeny. *Rep. USA Mar. Biol. Stat* **1970**, *17*, 1-179.

25. Cohen, D.M. Gadiformes: overview. In: Hubbs, C.; Moser, H.G.; Richards, W.J.; Cohen, D.M.; Fahay, M.; Kendall, A.W.; Jr Richardson SL. Ontogeny and Systematics of Fishes (Based on an International Symposium Dedicated to the Memory of Elbert Halvor Ahlstrom). Lawrence, Kansas, USA: American Society of Ichthyologists and Herpetologists. 2002.

26. Lloris, D.; Matallanas, J.; Oliver, P. Merluzas del mundo (familia Merlucciidae): catálago comentado e ilustrado de las merluzas conocidas / D. Lloris, J. Matallanas, P. OLiver. FAO Food and Nutrition Series. 2003.

27. Marshall, N.B.; Cohen, D.M. Order Anacanthini (Gadiformes): characters and synopsis of families. *In Fishes of the Western North Atlantic* **1973**, 479-495. New Haven: Sears Foundation for Marine Research, Yale University.

28. Okamura, O. Relationships of the suborder Macrouroidei and related groups, with comments on Merlucciidae and Steindachneria. In: Cohen DM. Papers on the Systematics of Gadiform Fishes. Natural History Museum of Los Angeles County. 1989.

29. Han, Z.Q.; Liu, M.H.; Liu, Q.; Zhai, H.; Xiao, S.H.; Gao, T.X. Chromosome-level genome assembly of burbot (*Lota lota*) provides insights into the evolutionary adaptations in freshwater. *Molecular Ecology Resources* **2021**, doi: 10.1111/1755-0998.13382

30. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology* **2019**, *1962*, 227-245. doi: 10.1007/978-1-4939-9173-0_14

31. Yang, T.Y.; Jiang, Y.L.; Han, Z.Q.; Meng, W. Comparative Transcriptome analysis of different tissues in burbot (*Lota lota*) based on RNA-seq. *Journal of Agricultural Biotechnology* **2020**, *28*, 291-301. doi: http://journal05.magtech.org.cn/Jwk_ny/CN/Y2020/V28/I2/291

32. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.D. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* **2011**, *29*, 644-652. doi: 10.1038/nbt.1883

33. Jaina, M.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **2013**, *41*, e121-e121. doi: 10.1093/nar/gkt263

34. Vincent, R.; Sébastien, H.; Delsuc, F.D.; Douzery, E.J.P.; Murphy, W.J. MACSE: Multiple alignment of coding sequences accounting for frame shifts and stop Ccodons. *Plos One* **2011**, *6*, e22594. doi: 10.1371/journal.pone.0022594

35. Roure, B.; Rodriguez-Ezpeleta, N.; Philippe, H. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evolutionary Biology* **2007**, *7*, S2. doi: 10.1186/1471-2148-7-S1-S2

36. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **2011**, *27*, 1164-1165. doi: 10.1093/bioinformatics/btr088

37. Alexandros S. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312-1313. doi: 10.1093/bioinformatics/btu033

38. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **2012**, *9(8)*, 772. doi: 10.1038/nmeth.2109

39. Li, H.; Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **2011**, *475*, 493-496. doi: 10.1038/nature10231

40. Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589-595. doi: http://doi.org/10.1093/bioinformatics/btp698

41. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987-2993. doi: 10.1093/bioinformatics/btr509