A chromosome-level genome assembly of Brachymystax lenok tsinlingensis provides new insights into salmonids evolution

Wenbo Zhu[1], Zhongkai Wang[2], Haorong Li[2], Hui Xiang[3], Ping Li[4], Lili Ni[1], Li Jiao[1], Yandong Ren[2], and Ping You[1]

[1]Shaanxi Normal University
[2]Northwestern Polytechnical University
[3]South China Normal University
[4]Shanghai Ocean University

May 19, 2021

Abstract

The salmonid-specific fourth vertebrate whole-genome duplication (Ss4R) occurred ˜80 million years ago in the ancestor of all salmonids and provides a unique opportunity to study the evolutionary history of the duplicated genome. Study of the genome of Brachymystax lenok tsinlingensis might be particularly insightful given that this is the only Brachymystax species with a published salmonid genome. Here, we present a high-quality chromosome-level genome assembly for B. l. tsinlingensis and found that the salmonids have a unique GC content and codon usage, have undergone a whole-genome duplication event and a burst of transposon-mediated repeat expansion, have a slower evolutionary rate, and possess specific expanded gene families and unique positively selected genes. Generally, the B. l. tsinlingensis genome could provide a valuable reference for the study of other salmonids as well as aid the conservation of this endangered species.

Running Head: Zhu et al. genome of *B. l. tsinlingensis*

A chromosome-level genome assembly of *Brachymystax lenok tsinlingensis* provides new insights into salmonids evolution

Wenbo Zhu[1,2,*], Zhongkai Wang[2,*], Haorong Li[2], Hui Xiang[3], Ping Li[4], Lili Ni[1], Li Jiao[1], Yandong Ren[2,+], and Ping You[1,+],

1. College of Life Sciences, Shaanxi Normal University, Xi'an, 710062, PR China.

2. School of Ecology and Environment, Northwestern Polytechnical University, Xi'an, 710072, PR China.

3. Guangdong Provincial Key Laboratory of Insect Developmental Biology and Applied Technology, Institute of Insect Science and Technology, School of Life Sciences, South China Normal University, Guangzhou, 510631, PR China.

4. Centre for Research on Environmental Ecology and Fish Nutrition (CREEFN) of the Ministry of Agriculture and Rural Affairs, Shanghai Ocean University,Shanghai 201306, PR China.

*These authors contributed equally to this work.

+These authors jointly directed this work. Correspondence should be addressed to P. Y. (youping@snnu.edu.cn), Y. R. (renyandong90@126.com).

Abstract

The salmonid-specific fourth vertebrate whole-genome duplication (Ss4R) occurred ˜80 million years ago in the ancestor of all salmonids and provides a unique opportunity to study the evolutionary history of the duplicated genome. Study of the genome of*Brachymystax lenok tsinlingensis* might be particularly insightful given that this is the only *Brachymystax* species with a published salmonid genome. Here, we present a high-quality chromosome-level genome assembly for*B. l. tsinlingensis* and found that the salmonids have a unique GC content and codon usage, have undergone a whole-genome duplication event and a burst of transposon-mediated repeat expansion, have a slower evolutionary rate, and possess specific expanded gene families and unique positively selected genes. Generally, the *B. l. tsinlingensis* genome could provide a valuable reference for the study of other salmonids as well as aid the conservation of this endangered species.

Keywords

Chromosome-level genome, salmonids, salmonid-specific fourth vertebrate whole-genome duplication, evolutionary rate, positively selected genes

Introduction

Whole-genome duplication (WGD) events have shaped the history of many evolutionary lineages; it is widely accepted that three rounds of WGD have occurred during vertebrate evolution: 1R and 2R are shared by the jawed vertebrates, and the third teleost-specific WGD (Ts3R) occurred at basally in the teleost radiation ˜320 million years ago (Mya) (Jaillon *et al.* 2004; Kasahara *et al.* 2007). However, the fourth salmonid-specific WGD (Ss4R) occurred in the common ancestor of salmonids ˜80 Mya after their divergence from Esociformes ˜125 Mya (Macqueen & Johnston 2014; Near *et al.* 2012). WGD events help species expand their genome size and evolve new characters; thus, studying the evolutionary fate of duplicated genomes, especially recently duplicated genomes, can provide valuable evolutionary insights.

*Brachymystax lenoktsinlingensis*is an endangered freshwater fish that is endemic to the middle part of the Qinling Mountains, especially the Heihe, Shitouhe, Xushui, and Taibaihe Rivers (Froese & Pauly 2014). Study of *B. l. tsinlingensis* could greatly aid our understanding of salmonid evolution given that it is the outgroup species of all known salmonids. However, only mitochondrial genome data have been published (Si *et al.* 2012; Yu & Kwak 2015); genome and transcriptome data of *B. l. tsinlingensis* are still lacking by comparison. In this study, we conducted the first chromosome-level genome assembly of *B. l. tsinlingensis* by combining Nanopore long reads, Hi-C data, and Illumina short reads. Comparative genomics analysis with other related species revealed that the salmonids have a unique GC content and codon usage, have undergone a common WGD event and a burst of transposon-mediated repeat expansion, have a slower evolution rate, and possess specific expanded genes and unique positively selected genes. Overall, these findings provide new insight into the evolutionary history of salmonids and will aid future studies of salmonid evolution.

Materials and Methods

Genome data generation

The muscle, heart, kidney, liver, and spleen were obtained from one adult, and all samples were stored in -80 for DNA and RNA extraction. Genomic DNA was extracted using a Qiagen DNeasy Blood & Tissue Kit. The quality of the extracted DNA was determined using an Agilent 2100 bioanalyzer (Agilent Technologies). The RNA of these tissues/organs was extracted using Trizol, and the quality of the extracted RNA was also assessed by an Agilent 2100 bioanalyzer (Agilent Technologies) and evaluated on a 1.5% agarose gel stained. Both DNA and RNA were used for the construction of different libraries and further sequencing. Genomic DNA was sequenced on a NovaSeq platform (Illumina) with a short-insert pair-end mode and Oxford Nanopore platform, respectively. RNA sequencing and Hi-C sequencing were both conducted on the NovaSeq platform (Illumina).

Quality control of raw sequencing reads

Because the sequencing data included Illumina short reads and Nanopore long reads, two different strategies were used to filter the reads. For Illumina data, all low-quality reads, adaptor sequences, and duplicated

reads produced by PCR were removed. For Nanopore reads, all reads with an average quality [?] 7 were retained for genome raw assembly.

Estimation of genome size

The genome size of *B. l. tsinlingensis* was estimated using the K-mer method with all filtered Illumina genome reads. The K-mer number was set to 17; the genome size was estimated using the total number of 17-mers divided by the peak 17-mer frequency. Thus, the genome size was estimated using the formula: Genome size = total 17-mer number / peak frequency (Liu *et al.* 2013).

De novo **genome assembly and Hi-C scaffolding**

All of the filtered Nanopore reads were used for the draft genome assembly in Nextdenovo software (v2.3.1; https://github.com/Nextomics/NextDenovo) with default parameters. The assembled genome was corrected using the filtered Illumina genome reads. BWA software (BWA-MEM module) (Li & Durbin 2009) was used to map short reads to the genome, and Pilon software (Walker *et al.* 2014) was used for error correction of the sequenced bases. To obtain the chromosome-level genome, the Hi-C reads were aligned to the assembled genome using Juicer software. The software 3D-DNA (Dudchenko *et al.* 2017) was used to cluster scaffolds into different clusters; in each group, the order of scaffolds was determined by the strength of interactions. Finally, all the possibilities of scaffold orientation and generated finely orientated scaffolds using a weighted directed acyclic graph.

Genome quality evaluation

After the genome assembly process, several genome quality evaluation methods were used. We used BUSCO (v2.0) software (Simao *et al.* 2015) to estimate the percentage of conserved orthologs in the assembled genome. The conserved gene sets of Eukaryota and Metazoa were used as a database and the complete BUSCO genes, fragment genes, and missing genes were detected. BWA (Li& Durbin 2009) and BLAT (Kent 2002) were used for the mapping ratio of NGS data and *de novo* assembled transcripts, respectively.

Repeats and transposable elements annotation

For repetitive sequence annotation, tandem repeats were annotated using Tandem Repeat Finder (TRF) (http://tandem.bu.edu/trf/trf.html, v4.10) (Benson 1999) with default parameters. For transposable elements (TEs) annotation, both RepeatProteinMask (RM-BLASTX) and RepeatMasker (open-4.0.7) (Bedell *et al.* 2000) were used. RepeatProteinMask software was used to search TEs in its protein database, and RepeatMasker software was used for *de novo*libraries and the Repbase library (zebrafish). The *de novo* repeat libraries were analyzed by RepeatModeler software with default parameters. The insertion time of each TE was calculated using$K$ */2r* (Bowen & McDonald 2001; SanMiguel *et al.* 1998).K represents the *kimura* value, which was extracted from the RepeatMasker analysis, and r represents the evolutionary rate acquired from the *r8s* analysis (Sanderson 2003).

Gene structure annotation and functional annotation

All of the TEs in the genome were masked and used for gene structure annotation. In this step, three different strategies, including *de novo* prediction, homolog-based prediction, and transcript-based prediction, were used. For *de novo* prediction, Augustus software (v3.3.3) (Stanke& Waack 2003) was used with default parameters. For homolog-based annotation, proteins of 10 species, including *Esox lucius* (GCF_-011004845.1) (Ishiguro *et al.* 2003),*Lepisosteus oculatus* (GCF_000242695.1) (Inoue *et al.* 2003), *Danio rerio* (GCF_000002035.6) (Howe *et al.* 2013), *Oncorhynchus tshawytscha* (GCF_002872995.1) (Christensen *et al.* 2018),*Oncorhynchus keta* (GCF_012931545.1), *Salmo salar*(GCF_000233375.1) (Davidson *et al.*2010), *Salmo trutta* (GCF_901001165.1), *Oncorhynchus nerka* (GCF_006149115.1), *Oncorhynchus mykiss*(GCF_-013265735.2), and *Oncorhynchus kisutch* (GCF_002021735.2), were downloaded from the NCBI database and aligned to the repeat-masked genome by tblastn (Altschul *et al.*1990) with an e-value of 10e-5. We then used Genewise software (Birney *et al.* 2004) to select the longest coding regions and/or the highest score at each gene locus. For transcript-based annotation, the RNA-seq reads were assembled into transcripts using

Bridger software (Chang*et al.* 2015), and the transcripts were mapped to the genome by BLAT software (v34) (more than 90% identity and coverage) (Kent 2002); PASA (Haas *et al.* 2003) was then used to link spliced alignments. Finally, EvidenceModeler (v1.1.1) (Haas *et al.* 2008) was used to integrate these results into the final gene set.

All of the predicted genes were used for functional annotation using the public protein database. InterProScan (v4.8) (Zdobnov& Apweiler 2001) was used to screen proteins against five databases (Pfam, release 24.057; ProDom, 2006.1; MART, release 6.059; PROSITE, release 20.52; PRINT, release 40.058). The Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt (Release 2011.6), non-redundant database (NR), and TrEMBL (Release 2011.6) databases were all used in the function annotation in BLAST software (v2.3.0) (Altschul *et al.*1990) with the e-value of 10e-5.

Genome synteny

Using the *B. l. tsinlingensis* genome as a reference assembly, 6 other salmonids (*O. mykiss* , *O. tshawytscha* , *S. trutta* , *O. kisutch* , *O. nerka* , and *O. keta*) were aligned to the *B. l. tsinlingensis* genome using LAST ("lastal" command in LAST with -P 5 -m100 -E 0.05; v802) (Kielbasa *et al.* 2011). The one-to-one aligned sequences were then selected and used for plotting with Circos (v0.69-6) (Krzywinski *et al.* 2009).

Phylogenetic inference

Reciprocal BLAST (Altschul *et al.*1990) and OrthoMCL software (Li& L. 2003) were used to determine the homology relationships among the protein sequences of 11 species (*B. l. tsinlingensis* , *D. rerio* , *O. mykiss* , *O. tshawytscha* , *S. trutta* ,*O. kisutch* , *O. nerka* , *S. salar* , *O. keta* ,*E. lucius* , and *L. oculatus*). After the orthologous genes were obtained, all of these genes in these species were connected into one super sequences in each species. MUSCLE (v3.8.31) (Edgar 2004a, b) was then used to align these sequences. Lastly, RAxML (v8.2.9) (Stamatakis& A. 2014) was used to reconstruct the phylogenetic relationships using different models (Protgammaauto/Gtrgamma) with 100 bootstrap replicates; *L. oculatus* was used as the outgroup species.

Molecular clock analysis

To estimate the divergence time among these 11 species, all of the 4d sites were extracted from the super sequences using in-house Perl scripts. The divergence times were analyzed by Markov chain Monte Carlo sampling with samples drawn every 2,000 steps and 100,000 samples, and the results were calibrated by the fossil records downloaded from the TIMEtree database (http://www.timetree.org).

Relative evolutionary rates

The evolutionary rate of these 11 species was calculated using the coding sequences (CDS) of all single-copy genes. Initially, all of these single-copy genes were connected into one super sequence and aligned using MUSCLE (v3.8.31) (Edgar 2004a, b). *L. oculatus* was used as the outgroup species; both LINTRE software (Takezaki *et al.* 1995) and MEGA (Tajima's relative rate test) (Kumar*et al.* 2016) were used for this analysis. In Tajima's test, the higher number of lineage-specific substitutions corresponds to a faster evolutionary rate. In the LINTRE method, the evolutionary rate of each species was checked using Z-statistics and the tpcv module.

Expansion and contraction of gene family

The expansion and contraction of gene family were determined by CAFE (-p 0.01, -r 1000, -s; v3.1) (De Bie *et al.* 2006) using the random birth-and-death (λ) model, and the probability of each gene family with observed sizes was calculated using 10,000 Monte Carlo simulations. Three results, including the phylogenetic relationships, divergence time, and fossil records, were used in this analysis. Both expanded gene families in *B. l. tsinlingensis* and the salmonid lineage were calculated, and these genes were used in Gene Ontology (GO)/ KEGG enrichment analysis (Beissbarth & Speed 2004; Huang da *et al.* 2009).

Positive selection

All sequences of these single-copy genes were aligned using MUSCLE software (Edgar 2004a, b) with default parameters. To estimate the lineage-specific evolutionary rate of each branch, the free model ("model = 1, NSsites = 0") in codeml was used. After we obtained a general evolutionary pattern of selective pressure along the lineages, the branch-site model was used. We used the likelihood ratio test and Chi-square test to calculate the *P-value* , and any positively selected genes with *P-values* less than 0.05 were retained.

WGD assessment

The distribution of synonymous substitutions per site (Ks) within paralogs was used to examine the most recent WGD event in salmonids. The protein sequences of all 11 species were aligned with BLAST v2.9.0 (e-value 1e-10) (Altschul *et al.*1990). When one gene and another gene were mutual best hits (excluding hits to themselves), they were identified as a paralog genes. Ks was calculated using the KaKs_calculator v2.0 (Wang *et al.* 2010) for each paralog. For comparison, we also plotted the Ks distribution between*B. l. tsinlingensis* and other species.

Hox gene cluster identification

Sequences of the known Hox genes were downloaded from the SwissProt database and used for Hox gene annotation in other species. Specifically, the downloaded sequences were aligned to the genome assembly using BLAST software (tblastn, v2.6.0) (Altschul *et al.* 1990) with an e-value of 1e-10. The lowest e-value of the annotation result was selected for the final annotation of each gene.

Results

Genome sequencing and chromosome-level genome assembly

We obtained a total of 714,523,240 sequence reads from the Illumina PE library, yielding a total sequence length of 100,539,703,234 bp (**Table S1**). K-mer analysis (K = 17) indicated that the genome size is 2.30 Gb with slightly repeats ratio (**Figure S1**). We also generated 100.54 Gb raw data using the Nanopore platform (111.95× coverage) for genome assembly: a data set of 13,970,063 reads with an N50 of 22,905 bp (**Table S2**). To construct the chromosome-level genome, 373.43 Gb Hi-C data were generated for scaffolding (**Table S3**). The final assembled *B. l. tsinlingensis*genome is summarized in **Table S4-S5** . In this assembly, 414 scaffolds/chromosomes (>100 bp) were included, with a total length of 2,031,709,341 bp. The N50 was 50.15 Mb, and the chromosome mounting ratio was 99.58% (**Figure 1A**). Although the published genomes of salmonids are mostly at the chromosome level, our genome is still comparable to other salmonids (**Table S6**). We also used different strategies to evaluate the quality of this genome assembly, including the assembled transcript mapping ratio from the RNA-seq data of 5 different tissues/organs (**Table S7-S9**), BUSCO results based on the Metazoan model sets and the Eukaryota model set (**Table S10**), Illumina short reads mapping ratio (**Table S11**), and the genomic synteny between *B. l. tsinlingensis* and 6 other salmonids (**Figure S2-S7**). All of these results indicated that this genome has a high level of accuracy, continuity, and connectivity.

Compared with other related species, the GC content of the genome and CDS of *B. l. tsinlingensis* were assessed. The genome GC content of all salmonids was similar and ranged from 43.03% to 43.55%, which is slightly higher compared with outgroup species such as *D. rerio* (36.60%), *E. lucius* (42.22%), and *L. oculatus*(39.59%) (**Figure S8**). The GC content of CDS of salmonids was also similar and ranged from 54.19% to 55.21%, which is slightly higher compared with outgroup species, such as *D. rerio*(49.85%), *E. lucius* (54.64%), and *L. oculatus* (53.25%) (**Figure S9**). These findings may be explained by the WGD event in the salmonids (salmonid-specific fourth vertebrate WGD, Ss4R) and suggest that other characters in salmonids, such as gene structure, TEs,*Hox* gene clusters, and others, may also be similar among salmonid fishes.

Genome annotation

From the assembled genome, the repeat sequences were identified in the genome of *B. l. tsinlingensis* . Repeat sequences accounted for 64.48% of the genome, and the DNA transposons (20.78%) were the most

abundant repeat type (**Table S12-S13**). For genome annotation, a total of 55,706 genes were predicted using different annotation methods, and the gene structure was similar to other published genomes of related species (**Figure S10**). The functional annotation results revealed that among these 55,706 protein-coding genes, homologous genes could be found in public databases for 90.14% (50,214) of the genes, which indicated that the gene structure annotation was robust (**Table S14**). The gene density, all types of repeat sequences, and GC density of the assembly are shown in **Figure 1B** . Owing to the Ss4R event, codon usage in salmonids was uniform (**Figure 1C**). In addition, the third position of the synonymous codon of salmonids was more likely to be G or C, which differs from other fishes (**Figure 1D**).

Salmonid-specific fourth vertebrate WGD

The last WGD for most teleosts occurred approximately 320 Mya (Jaillon *et al.* 2004; Kasahara *et al.* 2007); because of its basal occurrence in the teleost radiation, little is known about this WGD event. However, the salmonid-specific fourth vertebrate WGD (Ss4R) took place in the common ancestor of salmonids approximately 80 Mya (Macqueen& Johnston 2014; Near *et al.* 2012), which is the most recent WGD event in vertebrates. In addition, *B. l. tsinlingensis* is the most ancient species among known salmonids and thus particularly valuable for the study of salmonid evolution. To elucidate the history of Ss4R, we screened the paralogs of all species by McScan (Tang *et al.* 2008) and calculated the distribution of the rate of transversions on fourfold degenerate synonymous sites (4DTv). The peaks of all salmonids were concentrated near values of 4DTv at 0.2–0.3, which is the Ss4R event that occurred ˜80 Mya in all salmonids (Macqueen & Johnston 2014; Near *et al.* 2012). We also screened the orthologs with syntenic blocks between *B. l. tsinlingensis* and all other species separately. Furthermore, we calculated the 4DTv of the homologs in the other species, which showed that peaks of the 4DTv of the salmonids were around 0.7–0.9 (**Figure 2A**). Taken together, Ss4R occurred before the divergence of salmonids from Esociformes, and all salmonids shared the Ss4R event.

Because of the Ss4R event, the genome size of all salmonids (which ranges from 1.85 Gb to 2.97 Gb) is much larger than that of other related species (which ranges from 1.19 Gb to 1.68 Gb). The TEs in salmonids are much more abundant compared with other species, especially the tandem repeats, DNA elements, and LINE elements (**Figure 2B**). These three types of TEs may play important roles in WGD and contribute to the larger genome size of salmonids. In addition, the peaks of the TEs insertion time of salmonids were all around 40–80 Mya, which is earlier than Ss4R (**Figure 2C**). This indicates that the WGD and TE insertion events occurred in the common ancestor of salmonid fishes.

Hox genes of salmonids

Hox cluster organization provides a valuable marker for studying the effects of WGD in salmonids. For most teleosts with three rounds of WGD, the Hox gene clusters should number 7 or 8 (McArthur *et al.* 2003; Stellwag 1999). However, we found that the number of Hox gene clusters in salmonids was 13, including*HoξAαα* , *HoξAαβ* , *HoξAβα* , *HoξBαα* ,*HoξBαβ* , *HoξBβα* , *HoξBββ* , *Hoξα̂α* ,*Hoξα̂β* , *Hoξ˜βα* , *Hoξ˜ββ* , *HoξΔαα* , and*HoξΔαβ* (**Figure 3**). HoxAbβ appears to have been lost in the common ancestor of all salmonids because the most ancient salmonid species (*B. l. tsinlingensis*) does not possess the *HoξAββ*Hox cluster. With the exception of the assembly quality of salmonids, the 13 Hox clusters in salmonids were highly similar, which indicates that the Hox clusters were likely derived from a single common ancestor.

Comparative analysis with other species

Using the annotated genomes, we constructed orthologous gene clusters of*B. l. tsinlingensis,* with other salmonid fishes and other outgroup fishes. There were 31,296 gene families identified in these 11 species, and *B. l. tsinlingensis* contained 19,145 gene families. Because salmonid fishes have almost twice the number of genes compared with other species, the number of gene families is also twice that of other species (**Figure 4A**). In these 31,296 gene families, 11,343 gene families were salmonid-specific. Salmonids shared 16,840 (53.81%) gene families with *Esox lucius* , 14,783 (47.24%) with*Danio rerio* , and 13,491 (43.11%) with *Lepisosteus oculatus* (**Figure 4B**). A total of 1,216 1:1 single-copy orthologous genes were identified and used to construct phylogenetic trees, including a 4DTv tree, CDS tree, and protein tree. All of these trees

recovered the same phylogenetic topology with high confidence (**Figure S11-S13**). *B. l. tsinlingensis* and 8 other salmonid fishes formed a Salmonidae cluster, and *B. l. tsinlingensis* was the most ancient salmonid lineage. The divergence time between *E. lucius* and salmonid fishes was 132.9 Mya, and the divergence time between *B. l. tsinlingensis* and other salmonid fishes was 41.4 Mya (**Figure 4C**). Because of the Ss4R event in salmonids, we assessed the expanded gene families in salmonid lines and*B. l. tsinlingensis* , respectively. GO and KEGG enrichment analysis of salmonid expanded gene families indicated that most genes were related to protein binding, kinase activity, transferase activity, Toll signaling pathway, receptor activity, Glycolysis / Gluconeogenesis, citrate cycle, Pentose phosphate pathway, and Fatty acid biosynthesis, which indicates that metabolic processes in salmonids may differ from those in other fishes after the Ss4R event; the larger body size of salmonids might reflect differences in metabolic processes (**Table S15-S16**). The GO and KEGG enrichment analysis of *B. l. tsinlingensis* expanded gene families revealed that most genes were related to the regulation of apoptotic process, cell death, insulin receptor binding, lysozyme activity, biological regulation, ascorbate and aldarate metabolism, lipopolysaccharide biosynthesis, and mismatch repair, which indicates that the immune system of *B. l. tsinlingensis* might differ from that of other fishes (**Table S17-S18**). These unique expanded gene families can help us better understand the biology *B. l. tsinlingensis* and aid its conservation.

Evolutionary rate of salmonids

Studies of mutation frequencies over time have shown that species vary in their evolutionary rates. Thus, the evolutionary rate of *B. l. tsinlingensis* and other salmonids may be quite different from other fishes. The evolutionary rate of salmonids was quite similar and much slower compared with other outgroup species. However, the evolutionary rate of *B. l. tsinlingensis* was the fastest of all previously published evolutionary rates for salmonids (**Figure 4D; Table S19-S20**). These results indicate shown that after Ss4R, all of the salmonids occupied a stable environment and thus had a much slower evolutionary rate; however,*B. l. tsinlingensis* is still threatened and requires protection.

Positively selected genes

Species often face various several environmental pressures, which may explain the presence of several rapidly evolving, positively selected genes. Some specific gene loci may also undergo nucleotide substitutions. We determined positively selected genes in *B. l. tsinlingensis* and in the salmonid linage with *E. lucius* as the only outgroup species. *B. l. tsinlingensis* has the highest Ka/Ks value among these species, indicating that *B. l. tsinlingensis*possessed much more rapidly evolving genes (**Figure 4E**). This result is consistent with the findings above indicating that *B. l. tsinlingensis* has the fastest evolutionary rate among salmonids. There were 36 positively selected genes in*B. l. tsinlingensis* and 14 positively selected genes in salmonids. Four positively selected genes in *B. l. tsinlingensis* were related to the immune response, including*c1qb* ,*cebpb* ,*tnfa* , and *psme1* (**Table S21**). *c1qb* is associated with the brain immune system and social behavior (Ma *et al.* 2015), *cebpb* is a key immune-related gene that may be involved in sepsis (Xu *et al.* 2020),*tnfa*plays a vital role in the immune response by regulating several pathways that produce an immediate inflammatory reaction (Holbrook *et al.* 2019), and*psme1*may play an important role in the progression of Fanconi anemia to acute myeloid leukemia (Hou *et al.*2020). For the salmonid line, two positively selected genes were related to cell division and muscle development. *btg3* may play an important role in tumor suppression and is a key effector kinase in the cell cycle checkpoint response (Cheng*et al.*2013).*rapsn* has been shown to be associated with congenital myasthenic syndromes.*cld10*can form paracellular channels with ion selectivity; its variant causes anhidrosis and kidney damage (Joakim*et al.* 2017) (**Table S22**). Both muscle function and kidney function are important for the migration of salmonids, as salmonids need to be able to inhabit freshwater and seawater, which requires adaptation to both types of environments. The long-distance migration of salmonids also requires strong muscles. Both of these genes were positively selected in the salmonid line, which may be associated with the migratory habits of these species.

Discussion

B. l. tsinlingensis belongs to *Brachymystax* , which is the most ancient species of salmonid fish. Its genomic

resources not only provide new insights into its evolution but also supply basic data for future studies of salmonid evolution. We presented the first chromosome-level genome assembly of *B. l. tsinlingensis* with an N50 ˜50.15 Mb. Several different methods confirmed the high quality and accuracy of the assembled genome. Comparison of the genome of *B. l. tsinlingensis* with that of other salmonids revealed that the basic genome characters of salmonids were similar after the Ss4R event. Ss4R also caused an increase in TEs in salmonids and the number of genes, and several positively selected genes were specific to salmonids. The evolutionary rate of all salmonids was slower compared with other species. In addition, there is a pressing need to protect these endangered species. Habitat conservation for *B. l. tsinlingensis* is particularly important for ensuring the long-term survival and viability of *B. l. tsinlingensis* populations.

Acknowledgements

The authors would like to thank Professor Feng Wang (Yellow River Fisheries Research Institute, Chinese Academy of Fishery Science) for assistance with some materials.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215** , 403-410.

Bedell JA, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker.*Bioinformatics* **16** , 1040-1041.

Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20** , 1464-1465.

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

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* **14** , 988-995.

Bowen NJ, McDonald JF (2001) Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res* **11** , 1527-1540.

Chang Z, Li G, Liu J*, et al.*(2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol* **16** , 30.

Cheng YC, Lin TY, Shieh SY (2013) Candidate tumor suppressor BTG3 maintains genomic stability by promoting Lys63-linked ubiquitination and activation of the checkpoint kinase CHK1. *Proc Natl Acad Sci U S A* **110** , 5993-5998.

Christensen KA, Leong JS, Dionne S*, et al.* (2018) Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. *Plos One* **13** , e0195461.

Davidson WS, Koop BF, Jones SJ, Iturra... P (2010) Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biology* **11** , 403.

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.

Dudchenko O, Batra SS, Omer AD*, et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356** , 92-95.

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity.*Bmc Bioinformatics* **5** , 113.

Edgar RC (2004b) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* **32** , 1792-1797.

Froese R, Pauly D (2014) FishBase. World Wide Web electronic publication.

Haas BJ, Delcher AL, Mount SM*, et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research***31** , 5654-5666.

Haas BJ, Salzberg SL, Zhu W, Pertea... M (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.*Genome Biology* **9** , R7.

Holbrook J, Lara-Reyna S, Jarosz-Griffiths H, McDermott M (2019) Tumour necrosis factor signalling in health and disease [version 1; peer review: 2 approved].*F1000Res* **8** .

Hou H, Li D, Gao J*, et al.*(2020) Proteomic profiling and bioinformatics analysis identify key regulators during the process from fanconi anemia to acute myeloid leukemia. *Am J Transl Res* **12** , 1415-1427.

Howe K, Clark MD, Torroja CF*, et al.* (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496** , 498-503.

Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res***37** , 1-13.

Inoue JG, Miya M, Tsukamoto K, Nishida M (2003) Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". *Mol Phylogenet Evol* **26** , 110-120.

Ishiguro NB, Miya M, Nishida M (2003) Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii". *Mol Phylogenet Evol* **27** , 476-488.

Jaillon O, Aury JM, Brunet F*, et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature***431** , 946-957.

Joakim K, J?Rg P, Susanne M*, et al.* (2017) Altered paracellular cation permeability due to a rare CLDN10B variant causes anhidrosis and kidney damage. *PLOS Genetics* **13** , e1006897-.

Kasahara M, Naruse K, Sasaki S*, et al.* (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447** , 714-719.

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Res* **12** , 656-664.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison.*Genome Research* **21** , 487-493.

Krzywinski M, Schein J, Birol I*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* **19** , 1639-1645.

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology & Evolution* **33** , 1870.

Li, L. (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13** , 2178-2189.

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

Liu B, Shi Y, Yuan J*, et al.*(2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology***35** , 62-67.

Ma L, Piirainen S, Kulesskaya N, Rauvala H, Tian L (2015) Association of brain immune genes with social behavior of inbred mouse strains. *J Neuroinflammation***12** , 75.

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification.*Proc Biol Sci* **281** , 20132881.

McArthur AG, Hegelund T, Cox RL*, et al.* (2003) Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. *J Mol Evol* **57** , 200-211.

Near TJ, Eytan RI, Dornburg A*, et al.* (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci U S A* **109** , 13698-13703.

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

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* **20** , 43-45.

Si S, Wang Y, Xu G*, et al.*(2012) Complete mitochondrial genomes of two lenoks, Brachymystax lenok and Brachymystax lenok tsinlingensis. *Mitochondrial DNA***23** , 338-340.

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, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel.*Bioinformatics* **19 Suppl 2** , ii215-225.

Stellwag EJ (1999) Hox gene duplication in fish. *Semin Cell Dev Biol* **10** , 531-540.

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* **12** , 823-833.

Tang H, Bowers JE, Wang X*, et al.* (2008) Synteny and collinearity in plant genomes. *Science***320** , 486-488.

Walker BJ, Abeel T, Shea T*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One***9** , e112963.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8** , 77-80.

Xu C, Xu J, Lu L*, et al.*(2020) Identification of key genes and novel immune infiltration-associated biomarkers of sepsis. *Innate Immun***26** , 666-682.

Yu JN, Kwak M (2015) The complete mitochondrial genome of Brachymystax lenok tsinlingensis (Salmoninae, Salmonidae) and its intraspecific variation. *Gene* **573** , 246-253.

Zdobnov EM, Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17** , 847-848.

Data accessibility

All the raw genome sequencing data from different platforms, RNA-seq data, and the assembled genome during the current study are available in the National Centre for Biotechnology Information (NCBI) under BioProject accession number PRJNA713905.

Author contributions

P. Y. and W. Z. conceived and designed the investigation. W. Z., L. N., and L. J. performed field and laboratory work. Z. W. assembled the geome. H. L. performed the Hi-C scaffold. Y. R., H. X., W. Z., Z. W., and H. L. analyzed the data. P. L., P. Y., and W. Z. contributed materials and reagents. P. Y. and W. Z. wrote the paper. All the authors read and approved the final manuscript.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; RNA: Ribonucleic Acid; RNA-seq: RNA sequencing; BWA: Burrows-Wheeler Aligner; BLAST: Basic Local Alignment Search Tool; DNA: Deoxyribonucleic acid; KEGG: Kyoto Encyclopedia of Genes and Genomes; PCR: Polymerase Chain Reaction.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by grants from the National Natural Science Foundation of China (31872203) and Fundamental Research Funds for the Central Universities of Shaanxi normal university (2020TS051).

Tables and Figures

Figure 1. The assembled *B. l. tsinlingensis* genome and its codon usage characteristics. A: Heatmap of chromosome interactions in *B. l. tsinlingensis* . **B:** Circos graph of the genome characteristics. Shown from the outer circle to the inner ring are the gene distribution, tandem repeats (TRP), long tandem repeats (LTRs), short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), DNA elements (DNA), and genome GC content. **C:**Codon usage bias. Values of the codon bias index (CBI) on the frequency of guanine + cytosine at the synonymous third position of codons (GC3s) were determined using the nucleotide sequences of all predicted genes concatenated for individual species. **D:** The third position of the synonymous codon of these species.

Figure 2. Whole genome duplication events and the expanding genome. A: Ss4R event in *B. l. tsinlingensis* and other species. Solid lines are values for the 4DTv of paralogous genes in these 11 species; dotted lines are 4DTv of orthologs between *B. l. tsinlingensis* and other species. **B:** The genome composition of these species, including coding exons, tandem repeats, DNA elements, LINEs, SINEs, LTR elements, and others. **C:** The TE insertion times of these species.

Figure 3. *Hox clusters* in these species. Both Ts3R event and Ss4R event are shown in this figure.

Figure 4. Comparative genomics of *B. l. tsinlingensis* and other closely related species. A: Statistics of all orthologous/paralogous gene numbers in these species. **B:** A Venn diagram displaying the overlap in orthologous genes in salmonids and all 3 other fishes. **C:** The phylogenetic relationship and divergence time in these species. *L. oculatus* was used as an outgroup species. The red dot indicates that fossil record evidence was used here to adjust our divergence time results. **D:** The relative evolutionary rate of these species; the reference species was *B. l. tsinlingensis* . **E:** The Ka/Ks distribution of these species calculated by PAML.

Additional files

Table S1. The statistics of sequencing reads on Illumina platform.

Table S2. The statistics of sequencing reads on Nanopore platform.

Table S3. The statistics of Hi-C sequencing reads.

Table S4. The statistics of the polished genome and chromosome-level genome.

Table S5. Statistics of the assembled chromosome-level genome via 3D de novo assembly software.

Table S6. Comparison of all the released testudines genomes with our chromosome-level genome.

Table S7. The statistics of RNA sequencing reads on Illumina platform.

Table S8. The statistics of the assembled transcripts by Bridger of 5 organs/tissues.

Table S9. The statistics of the assembled transcripts mapping ratio on genome.

Table S10. The quality evaluation of assembled genome by BUSCO software.

Table S11. The statistics of the short reads mapping ratio on the assembled genome.

Table S12. The statistics of the annotated repeat sequences in our assembled genome.

Table S13. The statistics of the annotated repeat sequences in our assembled genome by de novo prediction.

Table S14. The functional annotation of the predicted protein-coding genes.

Table S15. GO enrichment of the expanded gene families in salmonids analyzed by CAFE.

Table S16. KEGG enrichment of the expanded gene families in salmonids analyzed by CAFE.

Table S17. GO enrichment of the expanded gene families in *B. l. tsinlingensis* genome analyzed by CAFE.

Table S18. KEGG enrichment of the expanded gene families in *B. l. tsinlingensis* genome analyzed by CAFE.

Table S19. Relative evolution rate among these species by LINTRE software.

Table S20. Relative evolution rate among these species by MEGA software.

Table S21. Positively selected genes in *B. l. tsinlingensis* .

Table S22. Positively selected genes in salmonids.

Figure S1. 17-mer analysis of the *B. l. tsinlingensis* genome.

Figure S2. Genomic synteny between *B. l. tsinlingensis* and*O. tshawytscha* .

Figure S3. Genomic synteny between *B. l. tsinlingensis* and*O. keta* .

Figure S4. Genomic synteny between *B. l. tsinlingensis* and*S. trutta* .

Figure S5. Genomic synteny between *B. l. tsinlingensis* and*O. nerka* .

Figure S6. Genomic synteny between *B. l. tsinlingensis* and*O. mykiss* .

Figure S7. Genomic synteny between *B. l. tsinlingensis* and*O. kisutch* .

Figure S8. The genome GC content of these 11 species.

Figure S9. The CDS GC content of these 11 species.

Figure S10. The distribution of gene stat in these species.

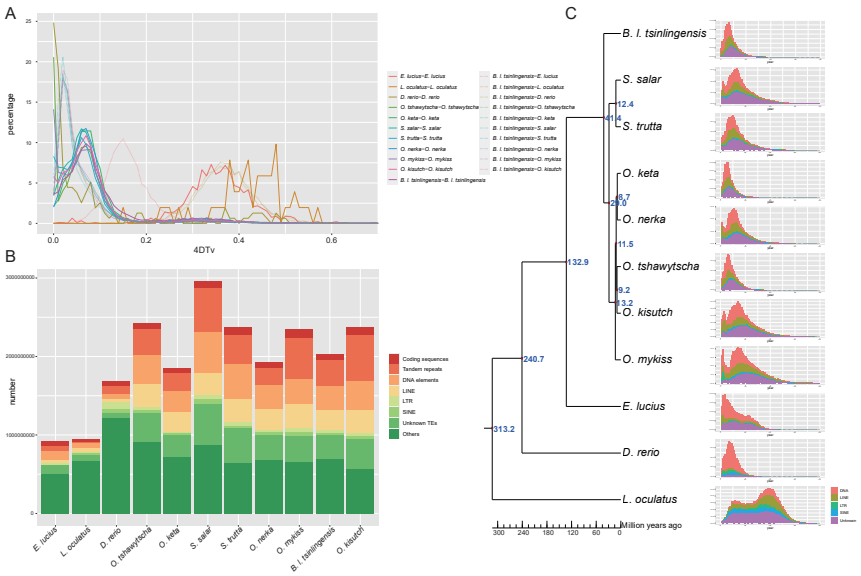
Figure S11. Phylogenetic relationship among the 14 species inferred by the 4dTV data of the single-copy genes.

Figure S12. Phylogenetic relationship among the 14 species inferred by the nucleotide acid sequences of the single-copy genes.

Figure S13. Phylogenetic relationship among the 14 species inferred by the amino acid sequences of the single-copy genes.

Hosted file

`Figure 1.pdf` available at https://authorea.com/users/414682/articles/522652-a-chromosome-level-genome-assembly-of-brachymystax-lenok-tsinlingensis-provides-new-insights-into-salmonids-evolution



Hosted file

Figure 3.pdf available at https://authorea.com/users/414682/articles/522652-a-chromosome-level-genome-assembly-of-brachymystax-lenok-tsinlingensis-provides-new-insights-into-salmonids-evolution

Hosted file

Figure 4.pdf available at https://authorea.com/users/414682/articles/522652-a-chromosome-level-genome-assembly-of-brachymystax-lenok-tsinlingensis-provides-new-insights-into-salmonids-evolution