Chromosome-level genome assembly for *hucho bleekeri* reveals the extraordinary tandem repeat proportions and its persistent population shrinkage

Xinmiao Zhang[1], Dongmei Xiong[1], Shenglong Jian[2], Yu Jiang[3], Lixin wang[1*]

[1]College of Animal Science and Technology, Northwest A&F University, Yangling, 712100, China

[2]Qinghai Provincial Fishery Technology Extension Center, Xining, China

[3]Center for Ruminant Genetics and Evolution, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China

[*]Corresponding author. E-mail: fisherwanglx@nwsuaf.edu.cn

Abstract

The Chuanshan hucho (*Hucho bleekeri*) is one of the largest freshwater salmonids, and located at the base of the Salmoninae phylogeny. Facing a serious risk of extinction, the Chuanshan hucho has attracted widespread concern and has been listed as national first-level protected animal in China. Currently, biological research and conservation policies of the Chuanshan hucho are in full progress, but the complete genome sequence is still unknown which hindered the advance of its conservation genomics. In this study, we provided the first chromosome-level genome for Chuanshan hucho, the final genome size was ~3.45 Gb with 44 chromosomes. The hucho genome contained 44.15% tandem repetitive sequence, which exceeded all sequenced salmonid fishes. According to the nuclear sequence, the divergence time of *Hucho* and *Brachymystax* was estimated as 12.29 Mya, and the effective population size of Chuanshan hucho declined from 1 Mya. The genomic resource in this article will promote the protection of Chuanshan hucho and evolutionary genetics studies of salmonids.

1 Introduction

The Chuanshan hucho (*Hucho bleekeri*) belonging to Salmoninae, is a glacial remnant species and endemic to China (Wang, et al. 2016). It has the lowest distribution latitude among all the fishes in *Hucho* genus, and mainly exists in the upper tributaries of the Yangtze River (Hu, et al. 2008). Hucho have a large body size and a ferocious habit, which occupied the top of the aquatic food chain (Wang, et al. 2016), and the late sexual maturation makes it more vulnerable to human activities and

34 ecological changes (Allendorf and Hard 2009; Sadovy de Mitcheson, et al. 2013).
35 Individuals of *H. bleekeri* were found sporadically and scattered in their distribution,
36 and had been protected since 1998, however, they continues to be illegally captured
37 by poachers for its high nutritional value and taste (Sung, et al. 1998). As the
38 continuous decline, *H. bleekeri* is listed as critically endangered (CR) level in the
39 International Union for Conservation of Nature (IUCN) Red List of Threatened
40 Species in 2012 (Song 2012). At present, *H. bleekeri* is under the highest level of
41 protection from the Chinese government (since 2021), and it is the only national
42 protected animal with "class I" among all distributed salmonid fishes.

43 Genomic data provide great opportunities for the progress on conservation and
44 evolution research of threatened animals (Kohn, et al. 2006; Allendorf, et al. 2010).
45 As sequencing prices have decreased, the genomes of many endangered fishes have
46 been published, such as *Trachidermus fasciatus* Heckel (Xue, et al. 2022),
47 *Myxocyprinus asiaticus* (Liu, et al. 2022), and *Brachymystax tsinlingensis*
48 (GCA_031305455.1), etc. Based on the genome data, the scientific and rational
49 conservation policies were formulated and further promoted the protection benefits of
50 the endangered animals (Brandies, et al. 2019; Waples, et al. 2020). However, lacking
51 of genome data of *H. bleekeri*, a blank was existence in the nuclear-based
52 phylogenetic research of Salmonidae, and also influenced the genomic studies such as
53 taxonomic studies based on SNPs (Chen, et al. 2020; Zhang, et al. 2020) and
54 researches on transcriptome (Chen, et al. 2021).

55 Due to the whole genome duplication and lineage-specific rediploidization in

56 salmonids, the genome of salmonids became complex and variable, increased the
57 difficulty of assembly (Robertson, et al. 2017; Dysin, et al. 2022). The circular
58 consensus sequencing (CCS) mode in single molecule real-time (SMRT) sequencing
59 of Pacific Biosciences can produce long and high fidelity (HiFi) reads, which
60 facilitates the genome assembly process and improves the quality (Wenger, et al.
61 2019). In this research, we utilized HiFi and Hi-C sequencing technology to construct
62 a high-quality genome for Chuanshan hucho, from the genome we found the existence
63 of high-proportioned tandem repetitive sequence. Additionally, we analyzed the
64 phylogenetic relationships of Chuanshan hucho and the demographic history of *H.*
65 *bleekeri* and *H. taimen*. Our work provides insights into the genome evolution of
66 salmonid fishes and the conservation of the endangered Chuanshan hucho.

67 # 2 Methods

68 ## 2.1 Sample collection and sequencing

69 Two wild individuals of *H. bleekeri* were collected from Taibai river (Baoji, Shannxi
70 province), one for genome assembly the other for resequencing, and an additional
71 wild individual of *H. taimen* was collected from Heilongjiang. All the experiments in
72 this study were performed in accordance with the guidelines of China Council on
73 Animal Care.

74 Genome DNA was extracted from kidney tissue with cetyltrimethylammonium
75 bromide (CATB) method (Stefanova, et al. 2013). The reads for genome size
76 estimation were generated on Illumina NovaSeq-6000 platform after the constructed
77 of 150 bp library. For genome assembly, a 15 kb library preparation was performed by
78 SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA), and

79 sequenced on the PacBio Sequel II platform (Pacific Biosciences, CA, USA) with two
80 cells of single molecule real time sequencing (SMRT). The CCS software v.6.0.0
81 (https://github.com/PacificBiosciences/ccs) was performed to integrate HiFi reads
82 with the parameters "--min-passes 3 --min-length 10 --min-rq 0.99". The Hi-C library
83 was constructed and sequenced on Illumina NovaSeq-6000 platform. The RNA
84 samples were extracted from fin， brain, gill, gonad, and eye, then sequenced on
85 Illumina NovaSeq-6000 platform. The two wild samples for demographic history
86 analysis were sequenced on DNBseq platform (BGI).

87 **2.2 Genome size estimation**

88 The raw Illumina paired-end reads of hucho were filtered using fastp version 0.20.0
89 (Chen, et al. 2018) with default parameters. The genome size estimation of hucho was
90 performed by GCE software version 1.0.2 (Liu, et al. 2013). A total of 58.89 Gb clean
91 reads were used to calculate the distribution of k-mer frequency using kmerfreq with
92 k=17. Then the gce was ran in both homozygous and heterzygous mode to evaluate
93 the genome size and heterozygosity. The genome heterozygous was calculated using
94 the formula "a[1/2]/k".

95 **2.3 Genome assembly**

96 The PacBio HiFi reads were assembled using hifiasm version 0.13-r308 (Cheng, et al.
97 2021) to obtain *de novo* contigs. After quality control by fastp (Chen, et al. 2018), the
98 clean Hi-C reads were used to integrate the contigs using Juicer version 1.6 (Durand,
99 Shamim, et al. 2016) and 3D-DNA version 201008 (Dudchenko, et al. 2017). First,
100 the Hi-C interaction matrix was generated using Juicer scripts with -s MboI parameter.
101 And then, the 3D-DNA with 0 rounds was used to link the contigs into chromosome

level based on the Hi-C interaction matrix. The initial assembly was manually corrected according to the Hi-C heatmap using Juicebox version 1.11.08 (Durand, Robinson, et al. 2016) and run the 3D-DNA again to generate the final assembly. The BUSCO (Benchmarking universal single-copy orthologues) version 3.0.2 (Simão, et al. 2015) was used to assess the accuracy and integrity of the genome with the gene database of ray-finned fishes (actinopterygii_odb9).

2.4 Repeat detection and gene annotation

To recognize repeat sequences of the hucho genome, we build a *de novo* library of transposable elements (TEs). First, we ran the RepeatModeler version 2.1 (Flynn, et al. 2020) with default parameters to constructed a repeat library; then an additional long terminal repeat (LTR) library of the hucho was built by LTR_FINDER version 1.07 (Xu and Wang 2007) and LTR_RETRIEVER version 2.9.6 (Ou and Jiang 2018); third the final repeat library was integrated by replacing the LTR elements of the first repeat library with the LTR library. To further classify the "Unknown" sequence, the BLAST version 2.14.0 (Ye, et al. 2006) was performed between the final repeat library and public repeat library (http://www.hrt.msu.edu/uploads/535/78637/Tpases020812.gz).

The TEs of the genome were described and masked by RepeatMasker version 4.1.2 (Tarailo-Graovac and Chen 2009) with parameters "-xsmall -s -cutoff 255 -frag 20000 –gff -engine ncbi". The tandem repeats (TRs) were detected through Tandem Repeats Finder (TRF) version 4.09 (Benson 1999) with parameters "2 7 7 80 10 50 500 -f -d -m". The TRs were further divided into short tandem repeats (STRs) and variable number tandem repeats (VNTRs) according to the length of repeat unit (STR:

125 unit number at 1 to 6; VNTR: unit number more than 6).

126 The gene annotation of the hucho genome was carried out using BRAKER2
127 pipeline version 2.1.5 (Hoff, et al. 2016; Hoff, et al. 2019). For transcriptome based
128 prediction, the RNAs were mapping to the assembled genome using STAR version
129 2.7.1a (Dobin, et al. 2013) with default parameters, then the RNA sequence data were
130 used to predict genes through auto BRAKER2 pipeline. For homology based
131 prediction, the protein sequence of six species including *Lepisosteus oculatus*
132 (GCF_000242695.1) (Braasch, et al. 2016), *Esox Lucius* (GCF_011004845.1)
133 (Ishiguro, et al. 2003), *Oncorhynchus mykiss* (GCF_013265735.2) (Gao, et al. 2021),
134 *Oncorhynchus nerka* (GCF_006149115.2) , *Salmo salar* (GCF_000233375.1) (Lien,
135 et al. 2016), and *Brachymystax tsinlingensis* (GCA_031305455.1) which represent the
136 sister clade or the ancestor species of hucho were download, and predicted through
137 BRAKER2 pipeline. The results from RNA and protein-based annotation were
138 combined using TSEBRA version 1.0.3 (Gabriel, et al. 2021) with default parameters
139 to generate the final gene prediction file. The gene expression level was measured by
140 transcript per million (TPM) which were calculated using Kallisto (Bray, et al. 2016)
141 with default parameters. To remove pseudogenes caused by tandem repetitive
142 sequence, we filtered the gene set based on the following criteria: 1) contain only one
143 exon; 2) TPM<1; 3) no matched BLAST results with *Brachymystax tsinlingensis*.

144 **2.5 Phylogenetic analysis**

145 Protein sequences of 10 species *Esox lucius*, *Coregonus* sp. *'balchen'*
146 (GCA_902810595.1) (Christensen, et al. 2020), *Thymallus thymallus*
147 (GCA_023634145.1), *Oncorhynchus mykiss*, *Oncorhynchus nerka*, *Salvelinus*

148 *namaycushs* (GCF_016432855.1) (Smith, et al. 2022), *Salmo salar*, *Salmo trutta*
149 (GCF_901001165.1) (Hansen, et al. 2021), *H. bleekeri*, and *Brachymystax*
150 *tsinlingensis* were used to construct phylogenetic tree. The single copy orthologous
151 proteins among the 10 species were recognized by Orthfinder version 2.3.12 (Emms
152 and Kelly 2019) with "msa" method. After alignment using MUSCLE version 3.8.31
153 (Edgar 2004), the species tree was calculated through RAxML version 8.2.12
154 (Stamatakis 2014) with 1000 bootstraps. The expansions and contractions of gene
155 family was estimated using CAFE5 (Mendes, et al. 2020).

156 The divergence time among all these species were calculated using MCMCtree
157 program of PAML software version 4.9d (Yang 2007). Four fossil records were
158 downloaded from TimeTree (Kumar, et al. 2022) and calibrated the results, the
159 Markov chain Monte Carlo (MCMC) was sampled with every 1000 step and conduct
160 10000 samples, the pre- burnin number was set at 20000.

161 **2.6 Demographic history estimation**

162 The pairwise sequentially Markovian coalescent (PSMC) method version 0.6.5-r67
163 (Li and Durbin 2011) was utilized to predict the effective population size. Besides the
164 two wild hucho sample, we used two lenok samples with highest sequencing depths
165 form Heilongjiang (*B.tumernsis*) and Shaanxi (*B.tsinlingensis*). Due to the different
166 sequencing depth, we used different strategies to generate the diploid consensus
167 sequences for huhco and lenok, -D 60 for hucho and –D 20 for lenok at vcfutils.pl
168 packages of bcftools version 1.12 (Danecek, et al. 2021). The psmc was run with
169 parameters "-N25 -t30 -r4 -p 8+4+25*2+4+6" to control the divergence at left of the
170 results.

3 Results

3.1 Genome size estimation and assembly

A total of 426685475 pair-end reads were obtained from the Illumina NovaSeq-6000 platform, with a total length of 64.00 Gb. The genome size estimated by GCE (k=17) was ~3.21 Gb with low heterozygous ratio 0.44% (Figure S1, Table S1). Based on the estimated genome size, the PacBio HiFi and Illumina Hi-C sequencing produced 76.48 Gb (23.8×) and 350.63 Gb (109.23×) raw reads separately for genome assembly, besides, 155.88 Gb resequencing data of the additional two wild sample (~22×, average) were generated.

After *de novo* genome assembly using the HiFi reads, a 3.48 Gb primary genome were generated which had 4373 high quality contigs (>2 Kb), and the contig N50 was 2.4 Mb (Table 1). After the Hi-C assisted genome assembly and scaffold deduplication, the contigs were anchored into 44 chromosome-level scaffolds, covered 95.26% total length of all scaffolds, and the scaffold N50 was 73.71 Mb (Figure 1; Table 1). Closed to the estimated size, the final hucho genome had a total size of 3.45 Gb with 41.99% GC ratio (Table 1). The BUSCO results showed that the assembled genome included 95.3% complete genes of actinopterygii_odb9 database, indicated the high completeness (Table S2).

3.2 Landscape of repeat sequences

The non-redundancy length of the total repeat sequence recognized by three annotation methods was 2.60 Gb, covered 74.21% of the hucho genome, which had the highest proportion of repetitive sequences among salmonids (Tables S3, S4 and S5). The interspersed repeats reported by the Repeatmasker and LTR_FINDER

194 composed 36.18% of the genome, which had the lowest proportion among salmonids,
195 but the total length in hucho genome was similar to others (Tables S3 and S4). The
196 proportion of retroelements in hucho genome was 13.11%, with 7.03% LINEs and
197 5.98% LTR elements (Tables S3). The DNA transposons composed 9.93% of the
198 genome, and the Tc1-IS630-Pogo family was the most abundant TE like other
199 salmonids and covered 9.38% of the hucho genome (Tables S3).

200 The tandem repetitive sequences covered 44.15% of the hucho genome, had the
201 highest proportion among salmonids (Figure 2; Tables S5). The lengths of STRs and
202 VNTRs were 45561499 bp and 1553872291, which accounted for 1.30% and 44.40%
203 of the genome respectively, and 59.12% of STRs had less than 1 Kb distance away
204 from the surrounding VNTRs (Figure 2). The distribution of TR was not uniformly,
205 we found 29.44% regions in hucho genome with low TR ratios (One-tenth of the
206 genome-wide scale; Figure 2), and some chromosomes had an enrichment of VNTR
207 at their ends (Figure 2). The distribution of STRs in hucho genome was relatively
208 scattered, only the top 200 STRs arranged by length exceeded the average length
209 (16620 bp) of the HiFi reads, had a total length of 15.64 Mb (Figure S2a).
210 Dinucleotide repeats accounted for 81.53% of STRs (Tables S6), which was higher
211 than the average level for vertebrates (Adams, et al. 2016).We identified 3337 VNTRs
212 with lengths over 100 Kb, which had a total length of 0.97 Gb and covered 28.20% of
213 the genome (Figure S2b). In addition, the presence of TRs affected GC content in
214 local regions of the genome, but the GC content of the TR region was 41.16%, similar
215 to the whole genome range (Figure 2).

216 **3.3 Gene annotation**

217 A total of 40.00 Gb RNA reads were generated from five tissues for RNA-based gene
218 annotation. After filtered out the pseudogenes, the BRAKER2 annotation pipeline
219 predicted 49052 genes, with an average length of 14576.6 bp. We generated 338448
220 exons, the average number of exon in gene was 6.90. After mapping with the NR
221 databases, only 79.02% of the predict genes were functionally annotated. The TR ratio
222 in intron regions was 22.74% which significantly lower than whole genome level
223 (Z-test; $P<0.01$). Based on the kernel density estimation, 161 gene hotspots were
224 identified (Figure S3). The results of NR database annotation showed that the top
225 three regions with the highest gene density were both histone gene family loci (chr8
226 and chr22). The result of BUSCO for the predicted genes showed that 88.1% were
227 complete BUSCOs (C:88.1%[S:43.3%,D:44.8%],F:4.3%,M:7.6%,n:4584).

228 **3.4 Analysis of phylogenetic and evolution of gene family**

229 We identified 36292 orthologous gene families between hucho and the other 9 species,
230 including 479 hucho specific gene families and 267 single copy orthologues. The
231 result of phylogenetic tree based on single copy orthologues revealed that the hucho
232 was a sister group to lenok, and the divergent time estimated by PAML was 12.29
233 Mya between this two species (Figure 3), which was closed to previous results based
234 on mitochondria (Kucinski and Fopp‐Bayat 2022). We generated 916 significant
235 positive expanded gene families in hucho genome, and the results of KEGG pathway
236 showed the enrichments the enrichment of motor and metabolic pathways such as
237 tight junction, adrenergic signaling in cardiomyocytes, glyoxylate and dicarboxylate
238 metabolism and regulation of actin cytoskeleton (Figure S4; Table S7), which was fits

239 with its ferocious hunting habits and large body size.

240 **3.5 Analysis of demographic history**

241 The geographical distribution pattern of *Brachymystax* is similar to that of *Hucho*, and
242 the competitive relationship is existed between them (Kaus, et al. 2019; Galland, et al.
243 2021). Due to the relatively larger body size and fiercer temperament, the hucho
244 dominated the niche and often preyed on lenok. In order to illustrate the trends of
245 historical effective population sizes (N_e) change of hucho and lenok in Shaanxi and
246 Heilongjiang, four samples of them were used for joint analysis. The result showed
247 that the history of four species traced back to different times respectively, ~9 Mya for
248 the two Heilongjiang sample *H. taimen* and *B.tumernsis*, and ~1 Mya for the two
249 Shaanxi sample *H. bleekeri* and *B.tsinlingensis* (Figure 4). The *H. taimen* had a
250 higher N_e level than *H. bleekeri* throughout the history (Figure 4). The curve
251 suggested that a large scale population expansion event occurred at ~2 Mya in both *H.*
252 *taimen* and *B.tumernsis*, and before ~250 Kya the N_e of *H. taimen* was greater than
253 the *B.tumernsis* (Figure 4). On the contrary, the *B.tsinlingensis* had a larger N_e than *H.*
254 *bleekeri* before ~150 Kya (Figure 4). In summary, the Heilongjiang populations had
255 higher N_e level than Shannxi populations since ~200 Kya, although the biomass of
256 hucho was currently smaller than that of lenok, there had been times in history with
257 the opposite case, which reflected the competition results between them.

258 **4.Discussion**

259 The combination of HiFi and Hi-C reads allowed us to achieve the chromosome-level
260 assembly of the largest salmonid genomes to date. The final size of hucho genome
261 was ~3.45 Gb, which exceeds that of the largest Atlantic salmon genome (~2.97 Gb)

262 (Lien, et al. 2016). The chromosome count of the hucho genome was 44, and
263 approached to some salmonids, such as lake trout (42), *B. tsinlingensis* (41), and
264 brown trout (40) (Table S8). The assembled genome recovered 95.20% complete
265 BUSCO genes, of which the percentage of duplicated BUSCOs (46.20%) was
266 relatively high among salmonids (Smith, et al. 2022), but similar to its sister group *B.*
267 *tsinlingensis*.

268 The genome assembly process and the outcome were affected by the ratio of
269 repetitive sequences (Treangen and Salzberg 2011; Tørresen, et al. 2017). The highly
270 accurate reads produced by circular consensus sequencing technology resulted in a
271 significant improvement in both the process and the quality of the assembly,
272 especially in the repetitive sequences region (Wenger, et al. 2019; Logsdon, et al.
273 2020). We found that the assembled hucho genome contains a high proportion of
274 repetitive sequences, and nearly half region of the genome were covered by tandem
275 repeat sequence, which may contribute to the increased genome size. We discovered
276 that most of the TR regions were covered and assembled by the HiFi reads, which
277 revealed the merit of long reads. However, there were some long TRs account for a
278 relatively large of the total genome length, and causing many gaps around them.
279 Generally, the number of gaps in the assembled hucho genome was similar to or even
280 smaller than other salmonid genomes, except for GCF_013265735.2, which indicated
281 that our assembly was not severely disturbed by TRs (Table S8).

282 The existence of long repetitive regions can affect the results of gene annotation,
283 leading to the appearance of false and redundant genes (Lang, et al. 2020). In our

284 preliminary gene annotation results, we found a mass of end-to-end pseudogenes in
285 long TR regions, which seriously interfered with our subsequent analysis. Despite the
286 presence of TRs in specific domains of some proteins (Simon and Hancock 2009), we
287 found that each of these pseudogenes had only one exon according to our manual
288 inspection, so we filtered out these single-exon genes located within the repetitive
289 sequence region based on expression (TPM<1). In parallel, we also blasted these
290 pseudogenes to demonstrate their absence in other species.

291 Salmonidae contains three subfamilies Coregoninae, Thymallinae, and
292 Salmoninae, but the evolutionary relationships remains uncertain (Dysin, et al. 2022).
293 Lineage-specific rediploidization, incomplete lineage sorting, hybridization, and
294 introgression resulted in the disordered messy phylogenetic relationship of
295 Salmonidae (Robertson, et al. 2017; Lecaudey, et al. 2018; Campbell, et al. 2020).
296 Some studies suggested that Thymallinae is the sister group of Salmoninae (McKay,
297 et al. 2004; Wang, et al. 2011; Crête-Lafrenière, et al. 2012), while some studies put
298 Thymallinae in the root of the Salmonidae phylogenetic tree (Horreo 2017; Wang, et
299 al. 2022). Based on the phylogenetic tree constructed by the orthologous genes, we
300 suggested that Coregoninae and Thymallinae had sister relationship which supported
301 by 100% bootstraps, and was consistent with other trees based on orthologues
302 (Macqueen and Johnston 2014; Robertson, et al. 2017), mitochondrial genome
303 (Campbell, et al. 2013; Shedko, et al. 2013), and ultra-conserved element (Campbell,
304 et al. 2020). The topology inner Salmoninae we inferred was consistent with other
305 studies, the *Hucho* and *Brachymystax* formed the ancient branch of the subfamily

306 salmoninae, and constituted the sister group to the genera *Salmo, Salvelinus and*
307 *Oncorhynchus* (Crespi and Fulton 2004; Kucinski and Fopp‐Bayat 2022). The
308 estimated divergence time based on nuclear genome was similar to the previous
309 researches based on mitochondrial genome (Crête-Lafrenière, et al. 2012; Horreo
310 2017; Kucinski and Fopp‐Bayat 2022). The divergence time between the *Hucho* and
311 *Brachymystax* was estimated at ~9.28 to ~17.96 Mya, which coincides with the
312 Miocene Climate Optimum (MCO) at 15-18 Mya (Böhme 2003; Sun and Zhang
313 2008).

314 All species of genera *Hucho* and *Brachymystax* were endangered and were
315 protected by Chinese government (Sung, et al. 1998), the results of PSMC revealed
316 that the effective populations size of four *Hucho* and *Brachymystax* species had
317 declined over 1 million years. The similar patterns of N_e curves were observed from
318 the species in same geographical distribution. The curves of two species *H. bleekeri*
319 and *B.tsinlingensis* in Shannxi had closed coalescent ending point, we speculated that
320 this two salmonids shared the same immigration route during the glacial period and
321 were retained in the low-latitude area at the interglacial period. The *H. bleekeri* had
322 lower N_e level than *H. taimen* in our calculation, and it was very rare in the current
323 resource survey in the Qinling Mountains, which had disappeared from most of the
324 original distribution areas (Shen, et al. 2006; Wang, et al. 2016). Consequently, more
325 effective measures should be taken to promote the conservation of *Hucho bleekeri*.

326 **Author contributions**

327 Xinmiao Zhang, and Lixin Wang conceived and supervised the project. Dongmei

Xiong, and Shenglong Jian the samples. Xinmiao Zhang, and Dongmei Xiong carried out the bioinformatic analyses. Xinmiao Zhang wrote the manuscript. Shenglong Jian, Yu Jiang and Lixin Wang revised the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

This study was financially supported by the National Natural Science Foundation of China (32273138) and the latitudinal project foundation of China Power Construction Group Beijing Survey, Design and Research Institute Co., Ltd. (K4050422184) . We thank the High-Performance Computing Platform of Northwest A&F University and the National Supercomputing Center in Xi'an and Hefei for providing computing resources.

342 **Reference**

343 Adams RH, Blackmon H, Reyes-Velasco J, Schield DR, Card DC, Andrew AL,
344 Waynewood N, Castoe TA. 2016. Microsatellite landscape evolutionary dynamics
345 across 450 million years of vertebrate genome evolution. Genome 59:295-310.
346 Allendorf FW, Hard JJ. 2009. Human-induced evolution caused by unnatural selection
347 through harvest of wild animals. Proc Natl Acad Sci U S A 106 Suppl
348 1:9987-9994.
349 Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of
350 conservation genetics. Nat Rev Genet 11:697-709.
351 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences.
352 Nucleic Acids Res 27:573-580.
353 Böhme M. 2003. The Miocene climatic optimum: evidence from ectothermic
354 vertebrates of Central Europe. Palaeogeogr, Palaeoclimatol, Palaeoecol
355 195:389-401.
356 Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A,
357 Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates
358 vertebrate evolution and facilitates human-teleost comparisons. Nat Genet
359 48:427-437.
360 Brandies P, Peel E, Hogg CJ, Belov K. 2019. The value of reference genomes in the
361 conservation of threatened species. Genes 10(11): 846.
362 Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic
363 RNA-seq quantification. Nat Biotechnol 34:525-527.
364 Campbell MA, Buser TJ, Alfaro ME, López JA. 2020. Addressing incomplete lineage
365 sorting and paralogy in the inference of uncertain salmonid phylogenetic
366 relationships. PeerJ 8:e9389.
367 Campbell MA, López JA, Sado T, Miya M. 2013. Pike and salmon as sister taxa:
368 detailed intraclade resolution and divergence time estimation of Esociformes +
369 Salmoniformes based on whole mitochondrial genome sequences. Gene
370 530:57-65.
371 Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ
372 preprocessor. Bioinformatics 34:i884-i890.
373 Chen Y, Yang H, Chen Y, Song M, Liu B, Song J, Liu X, Li H. 2021. Full-length
374 transcriptome sequencing and identification of immune-related genes in the
375 critically endangered *Hucho bleekeri*. Dev Comp Immunol 116:103934.
376 Chen Y, Yang H, Gong Q, Chen Y, Tu Q, Li H. 2020. Isolation and characterization of
377 34 SNP markers in *Hucho bleekeri*. Conserv Genet Resour 12:157-160.
378 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved *de novo*
379 assembly using phased assembly graphs with hifiasm. Nat Methods 18:170-175.
380 Christensen KA, Rondeau EB, Minkley DR, Sakhrani D, Biagi CA, Flores AM,
381 Withler RE, Pavey SA, Beacham TD, Godin T, et al. 2020. The sockeye salmon
382 genome, transcriptome, and analyses identifying population defining regions of
383 the genome. PLoS One 15:e0240935.
384 Crespi BJ, Fulton MJ. 2004. Molecular systematics of Salmonidae: combined nuclear

data yields a robust phylogeny. Mol Phylogenet Evol 31:658-679.
Crête-Lafrenière A, Weir LK, Bernatchez L. 2012. Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. PLoS One 7:e46662.
Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10(2): giab008.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.
Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92-95.
Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 3:99-101.
Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 3:95-98.
Dysin AP, Shcherbakov YS, Nikolaeva OA, Terletskii VP, Tyshchenko VI, Dementieva NV. 2022. Salmonidae genome: features, evolutionary and phylogenetic characteristics. Genes 13(12): 2221.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1-14.
Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451-9457.
Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. BMC Bioinf 22(1): 1-12.
Galland LM, Simmons JB, Jahner JP, Luzuriaga-Neira AR, Sloat MR, Chandra S, Hogan Z, Jensen OP, Parchman TL. 2021. Hierarchical genetic structure and implications for conservation of the world's largest salmonid, Hucho taimen. Sci Rep 11:20508.
Gao G, Magadan S, Waldbieser GC, Youngblood RC, Wheeler PA, Scheffler BE, Thorgaard GH, Palti Y. 2021. A long reads-based *de-novo* assembly of the genome of the Arlee homozygous line reveals chromosomal rearrangements in rainbow trout. G3 11(4): jkab052.
Hansen T, Fjelldal PG, Lien S, Smith M, Corton C, Oliver K, Skelton J, Betteridge E, Doulcan J, Fedrigo O, et al. 2021. The genome sequence of the brown trout, *Salmo trutta* Linnaeus 1758. Wellcome Open Res 6:108.
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1:

429 unsupervised RNA-seq-based genome annotation with GeneMark-ET and
430 AUGUSTUS. Bioinformatics 32:767-769.
431 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation
432 with BRAKER. Methods Mol Biol 1962:65-95.
433 Horreo JL. 2017. Revisiting the mitogenomic phylogeny of Salmoninae: new insights
434 thanks to recent sequencing advances. PeerJ 5:e3828.
435 Hu M, Wang Y, Cao L, Xiong BJEbof. 2008. Threatened fishes of the world: *Hucho*
436 *bleekeri* Kimura, 1934 (Salmonidae). Environ Biol Fishes 82:385-386.
437 Ishiguro NB, Miya M, Nishida M. 2003. Basal euteleostean relationships: a
438 mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii".
439 Mol Phylogenet Evol 27:476-488.
440 Kaus A, Michalski S, Hänfling B, Karthe D, Borchardt D, Durka W. 2019. Fish
441 conservation in the land of steppe and sky: Evolutionarily significant units of
442 threatened salmonid species in Mongolia mirror major river basins. Ecol Evol
443 9:3416-3433.
444 Kohn MH, Murphy WJ, Ostrander EA, Wayne RK. 2006. Genomics and conservation
445 genetics. Trends Ecol Evol 21:629-637.
446 Kucinski M, Fopp-Bayat DJJoAI. 2022. Phylogenetic analysis of *Brachymystax* and
447 *Hucho* genera-Summary on evolutionary status within the Salmoninae subfamily.
448 J Appl Ichthyol 38:403-411.
449 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G,
450 Hedges SB. 2022. TimeTree 5: an expanded resource for species divergence
451 times. Mol Biol Evol 39(8): msac174.
452 Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L, et al.
453 2020. Comparison of the two up-to-date sequencing technologies for genome
454 assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads
455 of Oxford Nanopore. Gigascience 9(12): giaa123.
456 Lecaudey LA, Schliewen UK, Osinov AG, Taylor EB, Bernatchez L, Weiss SJ. 2018.
457 Inferring phylogenetic structure, hybridization and divergence times within
458 Salmoninae (Teleostei: Salmonidae) using RAD-sequencing. Mol Phylogenet
459 Evol 124:82-99.
460 Li H, Durbin R. 2011. Inference of human population history from individual
461 whole-genome sequences. Nature 475:493-496.
462 Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS,
463 Minkley DR, Zimin A, et al. 2016. The Atlantic salmon genome provides insights
464 into rediploidization. Nature 533:200-205.
465 Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013.
466 Estimation of genomic characteristics by analyzing k-mer frequency in *de novo*
467 genome projects. arXiv.org arXiv:1308.2012.
468 Liu X, Zeng H, Wang C, Bo J, Gan X, Fang C, He S. 2022. Improved genome
469 assembly of Chinese sucker (*Myxocyprinus asiaticus*) provides insights into the
470 identification and characterization of pharyngeal teeth related maker genes in
471 Cyprinoidei. Water Biology and Security 1(3): 100049.
472 Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing

and its applications. Nat Rev Genet 21:597-614.
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.
McKay SJ, Trautner J, Smith MJ, Koop BF, Devlin RH. 2004. Evolution of duplicated growth hormone genes in autotetraploid salmonid fishes. Genome 47:714-723.
Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36:5516-5518.
Ou S, Jiang N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol 176:1410-1422.
Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol 18: 1-14.
Sadovy de Mitcheson Y, Craig MT, Bertoncini AA, Carpenter KE, Cheung WW, Choat JH, Cornish AS, Fennessy ST, Ferreira BP, Heemstra PCJF. 2013. Fishing groupers towards extinction: a global assessment of threats and extinction risks in a billion dollar fishery. Fish and fisheries 14:119-136.
Shedko SV, Miroshnichenko IL, Nemkova GA. 2013. Phylogeny of salmonids (Salmoniformes, Salmonidae) and molecular dating: analysis of mtDNA data. Genetika 49:718-734.
Shen Z, Tang W, Li K. 2006. The analysis of population dynamics of *Hucho bleekeri* in Markehe River, Qinghai Province. Reserv Fisheries 26:71-73 (In Chinese).
Simão 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.
Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol 10(6): 1-16.
Smith SR, Normandeau E, Djambazian H, Nawarathna PM, Berube P, Muir AM, Ragoussis J, Penney CM, Scribner KT, Luikart G, et al. 2022. A chromosome-anchored genome assembly for Lake Trout (*Salvelinus namaycush*). Mol Ecol Resour 22:679-694.
Song Z. 2012. *Hucho bleekeri*. The IUCN Red List of Threatened Species. Version 2014.3.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.
Stefanova P, Taseva M, Georgieva T, Gotcheva V, Angelov A. 2013. A modified CTAB method for DNA extraction from soybean and meat products. Biotechnol Biotechnol Equip 27:3803-3810.
Sun J, Zhang Z. 2008. Palynological evidence for the Mid-Miocene Climatic Optimum recorded in Cenozoic sediments of the Tian Shan Range, northwestern China. Global and Planetary Change 64:53-68.
Sung W, Peiqi Y, Yiyu C, Commission ESS. 1998. China Red data book of

endangered animals: Pisces. Science Press, Beijing, Hong Kong, New York, pp 32–34 (In Chinese).

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 4:1-4.10.

Tørresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, Walenz BP, Knight J, Ekholm JM, Peluso P, et al. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics 18:95.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36-46.

Wang K, Zhang SH, Wang DQ, Wu JM, Wang CY, Wei QW. 2016. Conservation genetics assessment and phylogenetic relationships of critically endangered *Hucho bleekeri* in China. J Appl Ichthyol 32:343-349.

Wang Y, Guo R, Li H, Zhang X, Du J, Song Z. 2011. The complete mitochondrial genome of the Sichuan taimen (*Hucho bleekeri*): repetitive sequences in the control region and phylogenetic implications for Salmonidae. Mar Genomics 4:221-228.

Wang Y, Xiong F, Song Z. 2022. Molecular Phylogeny and adaptive mitochondrial DNA evolution of Salmonids (Pisces: Salmonidae). Front Genet 13:903240.

Waples RS, Naish KA, Primmer CR. 2020. Conservation and management of salmon in the age of genomics. Annu Rev Anim Biosci 8:117-143.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155-1162.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-W268.

Xue DX, Xing TF, Liu JX. 2022. A high-quality chromosome-level genome of the endangered roughskin sculpin provides insights into its evolution and adaptation. Mol Ecol Resour 22:1892-1905.

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

Ye J, McGinnis S, Madden TL. 2006. BLAST: improvements for better sequence analysis. Nucleic Acids Res 34:W6-9.

Zhang Y, Luan P, Ren G, Hu G, Yin J. 2020. Estimating the inbreeding level and genetic relatedness in an isolated population of critically endangered Sichuan taimen (*Hucho Bleekeri*) using genome-wide SNP markers. Ecol Evol 10:1390-1400.

Table

TABLE 1 Summary statistics of the Chuanshan hucho (*Hucho bleekeri*) genome assembly.

Category	Length (bp)		Number	
	Contig	Scaffold	Contig	Scaffold
Max	25,857,850	143,121,144	-	-
Min	3,656	18,648	-	-
N50	2,400,826	73,705,547	354	19
N90	418,871	59,067,955	1680	39
Total	3,479,724,502	3,451,243,059	4373	505

Figures

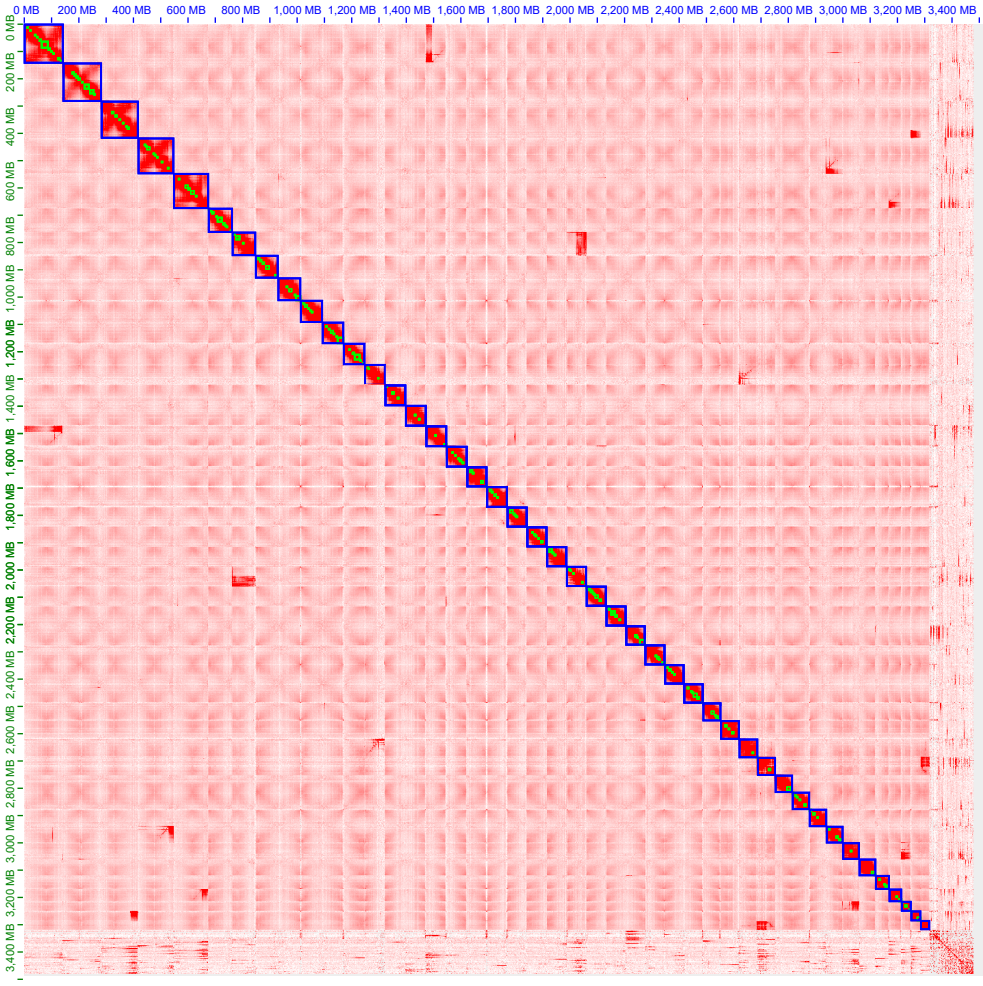

Figure 1 Hi-C interaction heatmap of *Hucho bleekeri* genome between 44 chromosomes and unmapped scaffolds. The intensity of red indicates the interaction strength. The 44 chromosome-level scaffolds are represented by blue outlines, and the preliminarily assembled contigs are shown by green outlines.

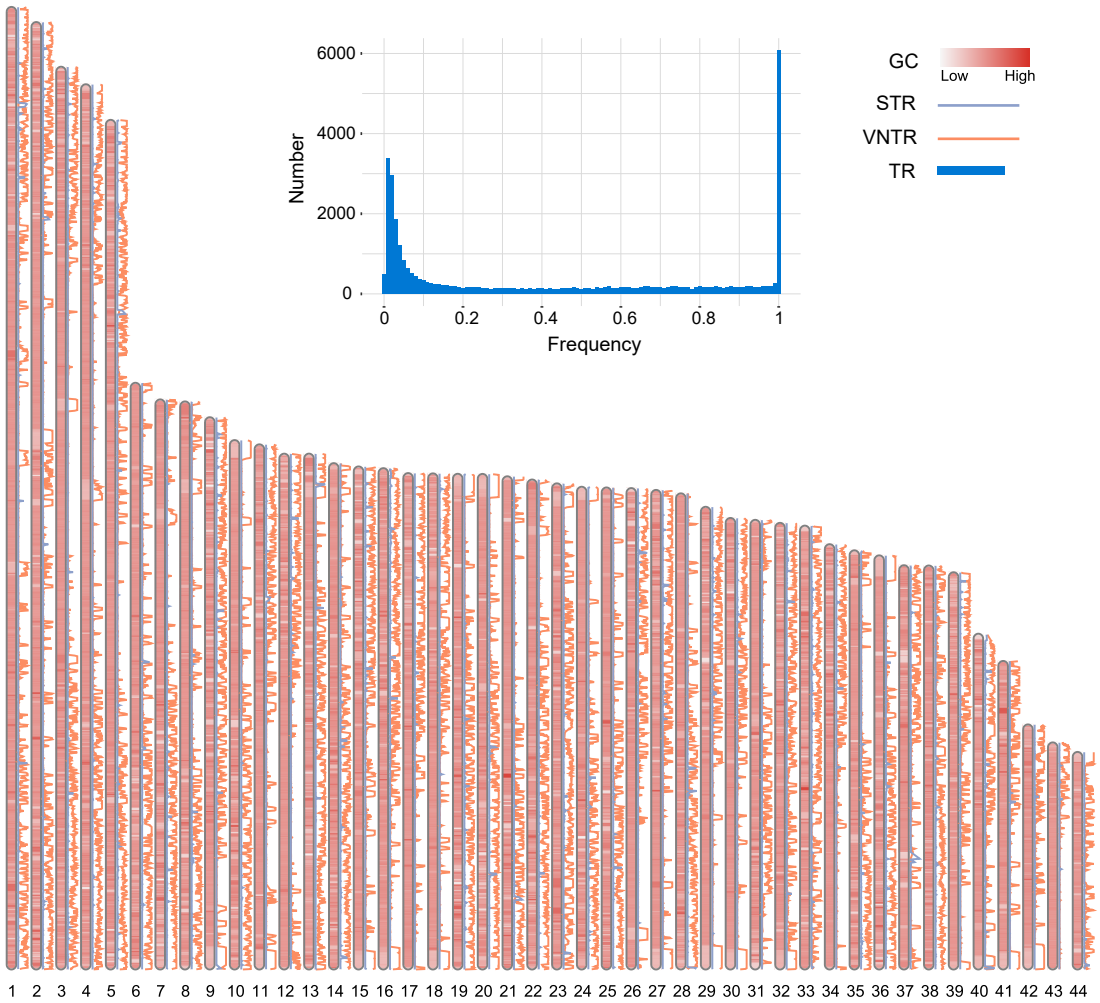

Figure 2 Visualization of genomic GC content, STR, and VNTR distribution based on 100 Kb windows. The bar graph reflects the TR frequency distribution in genome-wide 100 Kb windows.

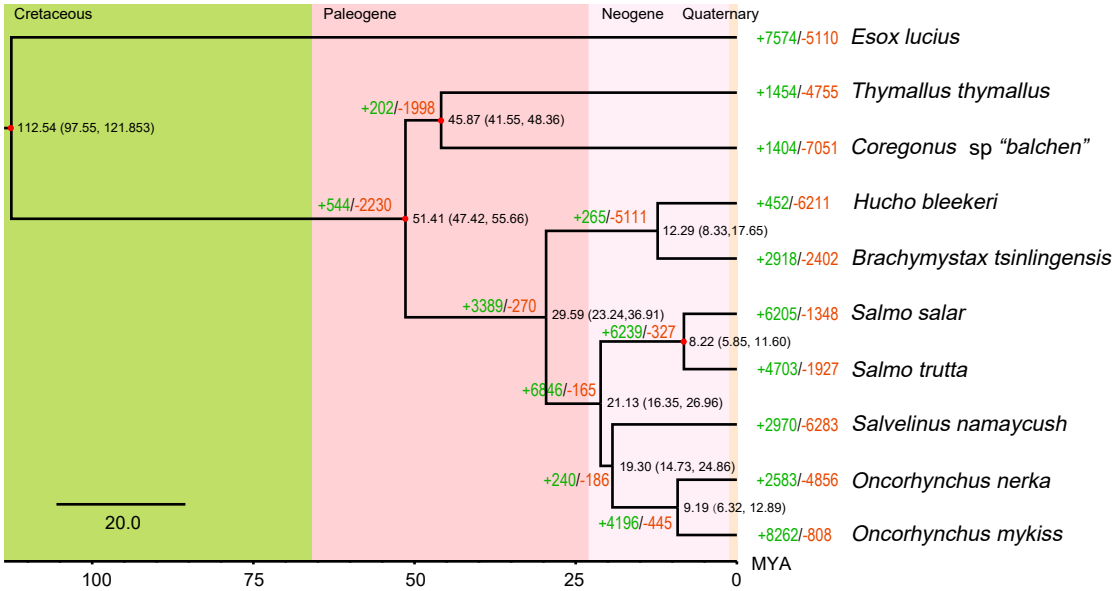

Figure 3 Maximum likelihood tree constructed based on single-copy orthologous genes from 9 salmonids and outgroup (*Esox lucius*). Fossil calibration points used by MCMC are marked in red. Black numbers next to each nodes are the divergence times (million years ago, MYA) and the 95% confidential intervals estimated by PAML. The green positive and red negative integers indicate the number of genes at expanded or contracted gene families at that node, respectively.

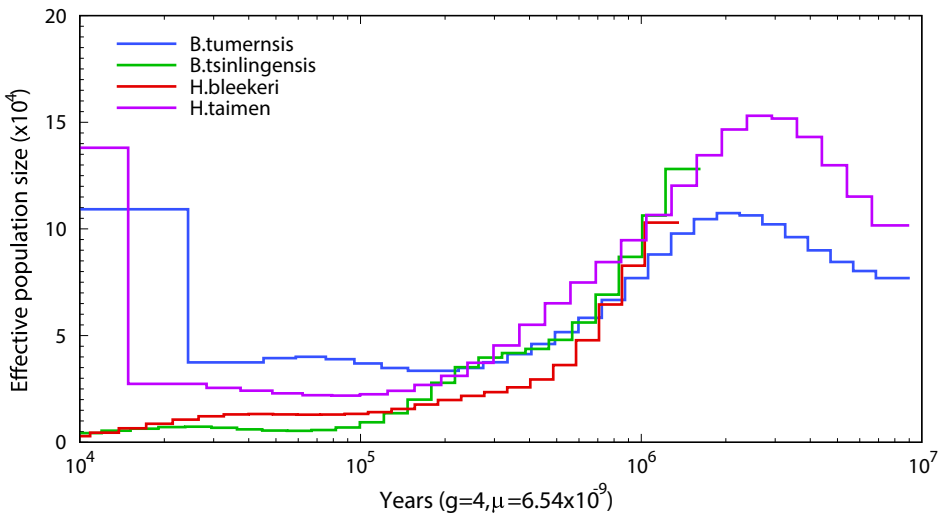

Figure 4 Demographic history of two Heilongjiang sample (*H. taimen* and *B.tumernsis*) and two Qinling sample (*H. bleekeri* and *B.tsinlingensis*). The generation (g) time was set at 4 years and the neutral mutation rate per generation (μ) was set at 6.54 × 10-9.

582 **Supplementary figure and table legend**

583 Supplementary Figure S1. Frequency distribution map in 17-mer.

584 Supplementary Figure S2. Statistics of tandem repeat unit length and total tandem
585 repeat sequence length of STRs (a) and VNTRs (b).

586 Supplementary Figure S3. Gene hotspots of the hucho genome.

587 Supplementary Figure S4. Visualization of the KEGG pathway annotation of the 1472
588 positive selected genes.

589 Supplementary Table S1. Genome size estimation table at k=17.

590 Supplementary Table S2. Summarized benchmarking in BUSCO notation.

591 Supplementary Table S3. Summary of TE in 10 salmonids (%).

592 Supplementary Table S4. Summary of TE in 10 salmonids (length).

593 Supplementary Table S5. Statistics of TR for 10 salmonids.

594 Supplementary Table S6. Summary of STRs' unit and total length.

595 Supplementary Table S7. The KEGG pathway annotation of the 1472 positive
596 selected genes.

597 Supplementary Table S8. Summary of genome assembly results of 10 salmonids.

598