Data Descriptor: The sequence and de novo assembly of *Oxygymnocypris stewartii* genome

Hai-Ping Liu¹,*, Shi-Jun Xiao¹,²,*, Nan Wu³,*, Di Wang³, Yan-Chao Liu¹, Chao-Wei Zhou¹, Qi-Yong Liu¹, Rui-Bin Yang⁶, Wen-Kai Jiang⁴, Qi-Qi Liang³, Wangjiu¹, Chi Zhang¹, Jun-Hua Gong¹, Xiao-Hui Yuan² & Zhen-Bo Mou¹,*

Animal genomes in the Qinghai-Tibetan Plateau provide valuable resources for scientists to understand the molecular mechanism of environmental adaptation. Tibetan fish species play essential roles in the local ecology; however, the genomic information for native fishes was still insufficient. *Oxygymnocypris stewartii*, belonging to Oxygymnocypris genus, Schizothoracinae subfamily, is a native fish in the Tibetan plateau living within the elevation from roughly 3,000 m to 4,200 m. In this report, PacBio and Illumina sequencing platform were used to generate ~385.3 Gb genomic sequencing data. A genome of about 1,849.2 Mb was obtained with a contig N50 length of 257.1 kb. More than 44.5% of the genome were identified as repetitive elements, and 46,400 protein-coding genes were annotated in the genome. The assembled genome can be used as a reference for future population genetic studies of *O. stewartii* and will improve our understanding of high altitude adaptation of fishes in the Qinghai-Tibetan Plateau.

| Design Type(s)          | sequence analysis objective • sequence annotation objective • sequence assembly objective • transcription profiling design |
|-------------------------|------------------------------------------------------------------------------------------------------------------|
| Measurement Type(s)     | whole genome sequencing • transcription profiling assay                                                              |
| Technology Type(s)      | DNA sequencing • RNA sequencing                                                                                         |
| Factor Type(s)          |                                                                                                                                 |
| Sample Characteristic(s)| Oxygymnocypris stewartii • muscle tissue • Tibetan Plateau • river • ovary • gut wall • kidney • adipose tissue • eye • swim bladder • skin of body • liver • heart • gill • brain |

¹Institute of Fisheries Science, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa 850002, China. ²School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China. ³Novogene Bioinformatics Institute, Beijing, China. ⁴College of Fishery, Huazhong Agricultural University, Wuhan, China.

*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to H.-P.L. (email: luihappying@163.com) or to Z.-B.M. (email: mouzhenbo@163.com)
Background & Summary

The Qinghai-Tibetan Plateau (QTP) is the largest and highest plateau in the world. The upshift of QTP has formed complex mountain systems in Southwest China and greatly reshaped the drainage at this area. The rapid alteration of topography in the QTP might act as significant barriers for gene flow of many species, leading to population isolations and initiating allopatric divergence and speciation. Genomes of fish species in the QTP provide valuable resources for scientists to understand the molecular mechanism of environmental adaptation. Although we have successfully obtained the reference genome of Glyptosternon maculatum, leading to the first high-quality fish genome in Tibet-plateau, the genome information of fish species in QTP is still lacking.

The schizothoracine fishes (Schizothoracinae subfamily, Cyprinidae family, Cypriniformes order), also known as “mountain carps”, which composed of approximately 100 species in 10–13 genera. They can be diagnosed by two lines of enlarged scales along both sides of the urogenital opening and anus. These fishes exhibit many unique traits that adapt to the extreme environment of the QTP. Therefore, this taxon provides an excellent opportunity for investigating high altitude adaptation of teleost fishes.

In this report, we provide the whole genome sequence of O. stewartii through the PacBio single molecule sequencing technique (SMRT). The availability of a fully sequenced and annotated genome is essential to support basic biological studies and will be helpful to the development of further protection strategies for this endangered species. Its whole genome sequence will also provide a foundation to explore the adaptive evolutionary processes of highland fishes, supplied as a starting point to study speciation mechanisms caused by the rapid rising of the QTP.

Methods

Sample collection and sequencing
A healthy female fish captured from Gongga Country, Lhasa, Tibet (Fig. 1a,b) was used for genome sequencing. Genomic DNA was isolated using Qiagen DNA purification kit (Qiagen, Valencia, CA, USA) from the white muscular tissue as in our previous studies. To generate enough read data for the genome assembly, both the PacBio SEQUEL and the Illumina HiSeq 4000 platform were used for the sequencing. Long reads generated from the PacBio platform were used for genome assembly, and the short but accurate reads from the Illumina platform were analyzed for genome survey and base level correction after the assembly. For the PacBio platform, genomic sequencing libraries were constructed according to the PacBio suggested protocol and 141.1 Gb long sequencing reads were obtained from 27 SMRT cells. A total of 140.7 Gb (coverage of 74.3×) subreads were obtained after removing adaptors in polymerase reads (Table 1). The subreads N50 and average lengths were 14.2 and 9.0 kb, respectively. For the Illumina HiSeq 4000 sequencing platform, one ug genomic DNA molecules were used for sequencing library construction. DNA molecules were fragmented, end-paired and ligated to the adaptor, which was further fractionated on agarose gels and purified by PCR amplification. To improve the representativeness of reads for the O. stewartii genome, 11 paired-end sequencing libraries were constructed with insert length of 250 bp according to Illumina’s protocol (Illumina, San Diego, CA, USA). Finally, a total of 145.4 Gb (coverage of 70.8×) short sequencing reads were generated. Reads with the adaptors and a quality value lower than 20 (corresponding to a 1% error rate) were filtered out. As a result, we obtained 144.3 Gb cleaned reads for the k-mer analysis and base correction of the genome (Table 1).

The individual used for the genomic sequencing was also used for the transcriptome sequencing, providing necessary gene expression data for the genome sequence annotation. Given that gene
expression exhibited clear tissue-specificity, 12 tissues, including skin, eye, swim bladder, muscle, brain, gill, heart, liver, gut, ovary, fat tissue and kidney were collected for the following transcriptome sequencing. As per the similar method in our previous study\textsuperscript{4}, RNA molecules were extracted using RNAiso Pure RNA Isolation Kit (Takara, Japan) for all samples, and DNase I treatment was performed to eliminate DNA contamination. After the quality assessment of the extracted RNAs using NanoVue Plus spectrophotometer (GE Healthcare, NJ, USA), RNA-seq libraries were constructed according to the protocol\textsuperscript{4} and were sequenced by Illumina HiSeq 4000 in paired-end 150 bp mode, resulting in a total of ~50 Gb transcriptome data. All genome and transcriptome sequencing data were summarized in Table 1.

De novo assembly of *Oxygymnocypris stewartii* genome

Genome size was estimated using Illumina sequencing data with the Kmer-based method\textsuperscript{16}. As per our previous study\textsuperscript{4}, we estimated the genome size of *O. stewartii* by the Kmer frequency distribution. Jellyfish (v2.1.3)\textsuperscript{17} was used to calculate the frequency of each Kmer from the short sequencing data (Table 2 and Fig. 2). As a result, we estimated the genome size of *O. stewartii* to be approximately 1,893.5 Mb.

The long reads generated from the PacBio SEQUEL platform were assembled into contigs using the FALCON package\textsuperscript{18} with default parameters. After the self-error correction step in the FALCON, we got

| Library types     | Insert size (bp) | Raw data (Gb) | Clean data (Gb) | Read length (bp) | Sequence coverage (X) |
|-------------------|------------------|---------------|-----------------|------------------|-----------------------|
| Illumina reads    | 250              | 145.4         | 144.3           | 150              | 76.21                 |
| Pacbio reads      | 20,000           | 141.1         | 140.7           | 13,287           | 74.31                 |
| RNA reads         | 250              | 98.8          | 94.76           | 150              | 50.04                 |
| Total             | —                | 385.3         | 379.76          | —                | 200.56                |

Table 1. Sequencing data used for the *Oxygymnocypris stewartii* genome assembly. Note that the coverage was calculated using the estimated genome size from the Kmer-based method.
104.9 Gb (55.4x coverage) of error-corrected pre-assembly reads. The assembly of the PacBio data alone resulted in a genome of 1,898.4 Mb with a contig N50 length of 240.3 kb. The assembled genomic sequences were further polished by two rounds of polishing with Quiver using the PacBio long reads. After that, another round of the genome-wide base-level correction was performed with the Illumina short sequencing data by Pilon. In the end, we obtained the final 1,849 Mb draft genome of *O. stewartii* with a contig N50 length of 257.1 kb (Table 3).

The completeness and the accuracy of the genome were evaluated by CEGMA, BUSCO and read mapping. The completeness of the genome assembly was assessed by the single copy orthologs (BUSCO, version 3.0) and CEGMA software. 94.2% complete and 3.6% partial of the 2,586 vertebrate BUSCO genes were identified in the final assembly. Using CEGMA, we revealed that 95.56% of the 248 core genes were evolutionarily conserved genes identified in the genome. Both BUSCO and CEGMA confirmed the completeness of the genome assembly. The accuracy of the genome was evaluated by the Illumina short read mapping with BWA and the transcript alignment with BLAT. More than 98.6% of the reads were aligned to the genome, and the insert length distribution exhibited a single peak that was

### Table 2. Statistics of 17-mer analysis for *Oxygymnocypris stewartii* genome. Note that all 17-mer sequences were extracted from paired-end clean reads that passed quality control (QC) from Next-generation sequencing libraries, and the frequency of each 17-mer was calculated and plotted in Fig. 2.

| Kmer | Kmer number | Peak depth | Genome size Mb | Used bases | Used reads | Coverage (X) |
|------|-------------|------------|----------------|------------|------------|--------------|
| 17   | 115,23,294,760 | 90         | 1,895.51       | 144,295,054,200 | 961,967,028 | 76.21        |

### Table 3. The statistics of length and number for the de novo assembled genome of *Oxygymnocypris stewartii*. Note that the length statistics of the genome assembly was based on the estimated genome size from the Kmer-based method.

| Statistics | Length (bp) | Number |
|------------|-------------|--------|
| Total      | 1,849,224,471 | 26,281 |
| Max        | 8,753,147   | —      |
| Number >= 2000 | 257,093 | 1,104 |
| N50        | 257,093     | 1,104  |
| N60        | 120,727     | 2,199  |
| N70        | 70,409      | 4,248  |
| N80        | 44,440      | 7,597  |
| N90        | 29,063      | 12,765 |
consistent with the experimental design. Meanwhile, the transcriptome was \textit{de novo} assembled by Trinity\textsuperscript{25}, and the transcripts were mapped to the genome assembly using BLAT\textsuperscript{24} with default parameters. We found that the alignment coverage (alignment length to transcript length) of expressed genes ranged from 96.44 to 99.95\% in the genome assembly.

| Type                | Repeat Size(bp) | percentage of genome (%) |
|---------------------|-----------------|--------------------------|
| TRF (Tendem Repeat Finder) | 151,169,214    | 8.17                     |
| RepeatMasker        | 788,753,932    | 42.65                    |
| RepeatProteinMask   | 103,814        | 0.01                     |
| Total               | 822,841,233    | 44.50                    |

Table 4. The annotation of repeated sequences in the \textit{Oxygymnocypris stewartii} genome using TRF, RepeatMasker, and RepeatProteinMask. Note that the total content was merged and redundancy was eliminated by each method.

| Type       | De novo+Repbase | TE Proteins | Combined TEs |
|------------|-----------------|-------------|--------------|
|            | Length (bp)     | % in Genome | Length (bp)  | % in Genome | Length (bp) | % in Genome |
| DNA        | 294,627,292     | 15.93       | 6,140        | 0.0003      | 294,628,980 | 15.93       |
| LINE       | 180,661,987     | 9.77        | 54,732       | 0.003       | 180,672,396 | 9.77        |
| SINE       | 10,828,447      | 0.59        | 0            | 0           | 10,828,447  | 0.59        |
| LTR        | 283,995,187     | 15.36       | 43,968       | 0.0024      | 284,039,155 | 15.36       |
| Satellite  | 35,364,895      | 1.91        | 0            | 0           | 35,364,895  | 1.91        |
| Simple_repeat | 37,479,121     | 2.03        | 0            | 0           | 37,479,121  | 2.03        |
| Unknown    | 25,680,794      | 1.39        | 0            | 0           | 25,680,794  | 1.39        |
| Total      | 788,753,932     | 42.65       | 103,914      | 0.0056      | 788,858,656 | 42.65       |

Table 5. Summary statistics of repeat annotation in \textit{Oxygymnocypris stewartii}. Note that \textit{De novo} + Repbase represent the result of RepeatMasker based on Repbase, RepeatModeler, RepeatScout, and LTR\_FINDER; TE proteins meant the result of RepeatProteinMask based on Repbase, and the Combined TEs refer to the combined results of \textit{De novo} + Repbase and TE proteins.

Figure 3. Distribution of the divergence rate of each type of repetitive element in \textit{Oxygymnocypris stewartii} genome. The divergence rate was calculated between the identified TE elements in the genome by the homology-based method and the consensus sequence in the Repbase.
To annotate repeat elements in the *O. stewartii* genome, both homologous comparison and *ab initio* prediction were applied. The similar annotation process in our previous work was employed. For *ab initio* repeat annotation, LTR_FINDER, RepeatScout, and RepeatModeler were used to construct a *de novo* repetitive element database, and RepeatMasker were used to annotate repeat elements with the database. Then, RepeatMasker and RepeatProteinMask were used for known repeat element types by searching against Repbase database. Tandem repeats were also *ab initio* predicted using TRF tool. A total of 822.84 Mb repetitive elements were identified in the *O. stewartii* genome by those repeat annotation processes, accounting for 44.50% of the whole genome (Tables 4 and 5 and Fig. 3).

For non-coding genes, 24,208 tRNAs were predicted using tRNAscan-SE, and 1,363 rRNA genes were annotated using BLASTN tool with an E-value of 1E-10 against human rRNA sequence. Small types of non-coding RNA were also annotated (Table 6).

| Type      | Number | Average length (bp) | Total length (bp) | % of genome |
|-----------|--------|---------------------|-------------------|-------------|
| miRNA     | 1,758  | 106.4               | 187,050           | 0.0101      |
| tRNA      | 24,208 | 75.45               | 1,826,526         | 0.0988      |
| rRNA      | 1,363  | 123.19              | 167,907           | 0.0091      |
| 18 S      | 112    | 294.73              | 33,010            | 0.0018      |
| 28 S      | 170    | 210.1               | 35,717            | 0.0019      |
| 5.8 S     | 19     | 103.42              | 1,965             | 0.0001      |
| 5 S       | 1,062  | 91.54               | 97,213            | 0.0053      |
| snRNA     | 923    | 132.36              | 122,168           | 0.0066      |
| CD-box    | 221    | 111.13              | 24,560            | 0.0013      |
| HACA-box  | 215    | 143.72              | 30,899            | 0.0017      |
| splicing  | 444    | 129.1               | 57,322            | 0.0031      |

Table 6. The number of the annotated non-coding RNA in the *Oxygymnocypris stewartii* genome.

| Methods/Tools | Gene Number | Average length (bp) | Exons number per gene |
|---------------|-------------|---------------------|-----------------------|
| Ab initio     |             |                     |                       |
| Augustus      | 101,732     | 7,592.54            | 981.37 188.14 1,568.10 | 5.22 |
| GlimmerHMM    | 223,822     | 7,337.30            | 534.39 154.34 2,762.67 | 3.46 |
| SNAP          | 198,963     | 10,915.28           | 755.07 150.73 2,534.08 | 5.01 |
| Genid         | 97,442      | 10,811.87           | 1,010.54 230.95 2,903.65 | 4.38 |
| Genescan      | 95,641      | 12,679.26           | 1,184.24 200.27 2,339.66 | 5.91 |
| Homolog       |             |                     |                       |
| Takifugu rubripes | 51,733   | 8,271.98            | 1,195.23 202.14 1,440.46 | 5.91 |
| G. stellatus  | 70,092      | 6,437.54            | 1,162.26 217.66 1,220.15 | 5.34 |
| Danio rerio   | 63,215      | 8,666.61            | 1,261.85 206.17 1,407.08 | 6.12 |
| Cyprinus carpio | 78,104    | 6,467.89            | 1,176.98 227.61 1,248.52 | 5.17 |
| Mus musculus  | 44,944      | 9,259.53            | 1,202.88 189.77 1,509.12 | 6.34 |
| A. punctatus  | 59,212      | 8,747.74            | 1,268.06 205.62 1,447.61 | 6.17 |
| Sinocypris griffithi | 70,380   | 7,956.29            | 1,204.04 205.73 1,391.50 | 5.85 |
| Homo sapiens  | 46,989      | 9,041.85            | 1,176.04 189.54 1,511.28 | 6.2  |
| RNA-seq       |             |                     |                       |
| Cufflinks     | 93,109      | 21,118.90           | 3,436.98 357.39 2,051.98 | 8.62 |
| PASA          | 140,045     | 10,537.33           | 1,152.91 165.15 1,569.00 | 6.98 |
| EVM           | 101,031     | 8,674.09            | 1,018.48 183.65 1,684.16 | 5.55 |
| PASA-update   | 100,450     | 8,739.14            | 1,026.34 184.60 1,691.47 | 5.56 |
| Final set     | 46,400      | 13,348.16           | 1,438.34 171.04 1,607.39 | 8.41 |

Table 7. The statistics of gene models of protein-coding genes annotated in the *Oxygymnocypris stewartii* genome. Note that: CDS refers to coding sequence; GlimmerHMM was a new gene finder based on a Generalized Hidden Markov Model (GHMM); SNAP refers to Semi-HMM-based Nucleic Acid Parser; EVM refers to Evidence modeler.

Repetitive element and non-coding gene annotation in the *O. stewartii* genome
To annotate repeat elements in the *O. stewartii* genome, both homologous comparison and *ab initio* prediction were applied. The similar annotation process in our previous work was employed. For *ab initio* repeat annotation, LTR_FINDER, RepeatScout, and RepeatModeler were used to construct a *de novo* repetitive element database, and the RepeatMasker (http://repeatmasker.org/RMDownload.html) were used to annotate repeat elements with the database. Then, RepeatMasker and RepeatProteinMask were used for known repeat element types by searching against Repbase database. Tandem repeats were also *ab initio* predicted using TRF tool. A total of 822.84 Mb repetitive elements were identified in the *O. stewartii* genome by those repeat annotation processes, accounting for 44.50% of the whole genome (Tables 4 and 5 and Fig. 3).

For non-coding genes, 24,208 tRNAs were predicted using tRNAscan-SE, and 1,363 rRNA genes were annotated using BLASTN tool with an E-value of 1E-10 against human rRNA sequence. Small
nuclear and nucleolar RNAs in the *O. stewartii* genome were also annotated by the infernal tool \(^{33}\) using Rfam database \(^{34}\) (Table 6).

### Protein-coding gene prediction and functional annotation

The gene model prediction method in our previous study \(^{4}\) was applied to the protein-coding gene annotation in the *O. stewartii* genome. We merged the evidence of the gene prediction from multiple methods, including homolog based, *ab initio* and RNA-seq based annotations. The protein and coding sequences were obtained from the Ensembl database \(^{35}\) for the following species, including human (*Homo sapiens*, GCF_000001405.37), mouse (*Mus musculus*, GCF_000001635.26), zebrafish (*Danio rerio* var, GCF_000180615.1), tiger puffer (*Takifugu rubripes*, GCF_000180615.1), channel catfish (*Ictalurus punctatus*, GCF_001660625.1), Sinocyclocheilus graham (GCF_001515645.1) and grass carp \(^{36}\) (*Ctenopharyngodon idellus*). The protein sequences were aligned against the *O. stewartii* genome using TBLASTN \(^{37}\) search with parameters of e-value 1e-5. After filtering low-quality records, the gene structure was predicted by GeneWise \(^{38}\) (referred to “Homology” in Table 7). Secondly, transcripts assembled from twelve tissues RNA-Seq data were aligned against the *O. stewartii* genome using Program to Assemble Spliced Alignment (PASA) \(^{39}\) (referred to “PASA” in Table 7). Augustus \(^{40}\), GeneID \(^{41}\), GeneScan \(^{42}\), GlimmerHMM \(^{43}\), and SNAP \(^{44}\) were used for *ab initio* prediction with the optimized parameters that trained using high-quality proteins that derived from the PASA gene models. RNA-seq reads were also aligned to the *O. stewartii* genome directly using TopHat \(^{45}\) v2.0.9, and the gene models were constructed by Cufflinks \(^{46}\) v2.2.1 (referred to “Cufflinks” in Table 7). Finally, EvidenceModeler \(^{39}\) was applied to combine all gene models that were predicted by various methods with the identical weights with our previous work \(^{4}\). Untranslated regions (UTRs) and alternative splicing variations were annotated using PASA2 \(^{39}\) (referred to “PASA-update” in Table 7). Finally, 46,400 protein-coding genes with a mean of 8.41 exons per gene (Table 7) were annotated in the *O. stewartii* genome. The statistics of gene models, including lengths of a gene, CDS, intron, and exon in *O. stewartii* were comparable to those for close-related species (Table 8 and Fig. 4). Public biological function databases of SwissProt \(^{47}\), InterPro \(^{48}\), NR from NCBI and Kyoto Encyclopedia of Genes and Genomes (KEGG) \(^{49}\) were used for the functional annotation of the predicted genes. BLASTX utility \(^{32}\) were used for the homolog search with an E-value threshold of 1E-5. InterPro database \(^{48}\) was used to predict protein function based on the conserved protein domains by InterproScan tool \(^{50}\). A total of 45,991 genes (99.1%) were successfully annotated by at least one public database. (Table 9 and Fig. 5).

### Code Availability

The sequence data were generated using the software provided by the sequencing platform manufacturer and the sequencing data were processed with commands with the guidance from the public software that is cited in the manuscript. No custom computer codes were generated in this work.

### Data Records

All PacBio long-read sequencing data and Illumina short-read sequencing data have been deposited to NCBI Sequence Read Archive (SRA) (Data Citation 1). The transcriptome data are available through the NCBI SRA (Data Citation 2). The assembled genome version is available at GenBank (Data Citation 3). The annotation gff3 file of the assembled genome is available at Figshare (Data Citation 4).

---

Table 8. The comparison of the gene models annotated from the *Oxygymnocypris stewartii* genome and other teleosts.

| Species                        | Gene Number | Average length (bp) | Exons number per gene |
|--------------------------------|-------------|---------------------|-----------------------|
|                                |             | transcript         | CDS                   | Exon       | Intron     |                                                                 |
| *Oxygymnocypris stewartii*     | 46,400      | 13,348.16          | 1438.34               | 171.04     | 1,607.39   | 8.41                                                          |
| *Ctenopharyngodon idellus*     | 32,811      | 10444.53           | 1384.98               | 180.99     | 1,361.89   | 7.63                                                          |
| *Homo sapiens*                 | 19,805      | 43772.47           | 1457.89               | 171.22     | 5631.04    | 8.51                                                          |
| *Mus musculus*                 | 22,278      | 37435.55           | 1600.64               | 179.14     | 4516.11    | 8.93                                                          |
| *Sinocyclocheilus grahami*     | 45,899      | 16243.9            | 1585.31               | 171.68     | 1780.2     | 9.23                                                          |
| *Takifugu rubripes*            | 21,317      | 8334.84            | 1699.01               | 165.45     | 715.91     | 10.27                                                         |
| *Danio rerio*                  | 25,619      | 25207.59           | 1642.64               | 174.39     | 2798.97    | 9.42                                                          |
| *Cyprinus carpio*              | 49,264      | 11780.68           | 1260.28               | 163.96     | 1573.34    | 7.69                                                          |
| *Ictalurus punctatus*          | 22,966      | 17866.19           | 1760.81               | 170.99     | 1732.19    | 10.3                                                          |

---
Technical Validation

RNA integrity

The transcriptomes for twelve tissues from three fish individuals were sequenced. Before constructing RNA-Seq libraries, the concentration and quality of total RNA were evaluated using NanoVue Plus spectrophotometer (GE Healthcare, NJ, USA). The total amount of RNA, RNA integrity and rRNA ratio were used to estimate the quality, content and degradation level of RNA samples. In the present study, RNAs samples with a total RNA amount $\geq 10 \, \mu g$, RNA integrity number $\geq 8$, and rRNA ratio $\geq 1.5$ were finally subjected to construct the sequencing library.

| Database | Annotated Num | Annotated Percent (%) |
|----------|---------------|-----------------------|
| NR       | 45,976        | 99.1                  |
| Swiss-Prot| 43,115        | 92.9                  |
| KEGG     | 39,302        | 84.7                  |
| InterPro |               |                       |
| All      | 43,183        | 93.1                  |
| Pfam     | 38,742        | 83.5                  |
| GO       | 31,811        | 68.6                  |
| Annotated| 45,991        | 99.1                  |
| Total    | 46,400        | -                     |

Table 9. The number of genes with homology or functional classification for *Oxygynocypris stewartii*.

Figure 4. Comparisons of the prediction gene models in the *Oxygynocypris stewartii* genome to other species. (a) CDS length distribution and comparison with other species. (b) Exon length distribution and comparison with other species. (c) Exon number distribution and comparison with other species. (d) Gene length distribution and comparison with other species. (e) Intron length distribution and comparison with other species.
Quality filtering of Illumina sequencing raw reads

The raw sequencing reads generated from the Illumina platform were rigorously cleaned by the following procedures as in the previous study. Firstly, adaptors in the reads were filtered out; secondly, reads with more than 10% of N bases were filtered out; thirdly, reads with more than 50% of the low-quality bases (phred quality score < 5) were filtered out. If any end pair was classified as low quality, both pairs were discarded. The initially generated raw sequencing reads were also evaluated for quality distribution, GC content distribution, base composition, average quality score at each position and other metrics.

References

1. Liu, J.-Q., Wang, Y.-J., Wang, A.-L., Hideaki, O. & Abbott, R. J. Radiation and diversification within the Ligularia—Crematodorum—Parasenecio complex (Asteraceae) triggered by uplift of the Qinghai-Tibetan Plateau. *Molecular Phylogenetics and Evolution* 38, 31–49 (2006).
2. Pan, G et al. Tectonic evolution of the Qinghai-Tibet Plateau. *Journal of Asian Earth Sciences* 53, 3–14 (2012).
3. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution* 16, 351–358 (2001).
4. Liu, H. et al. Draft genome of Glyptosternon maculatum, an endemic fish from Tibet Plateau. *GigaScience* 7, giy104–giy104 (2018).
5. Mirza, M. R. A contribution to the systematics of the Schizothoracine fishes (Pisces: Cyprinidae) with the description of three new tribes. *Pakistan Journal of Zoology* 23, 339–341 (1991).
6. Cao, W., Chen, Y., Wu, Y. & Zhu, S. Origin and evolution of Schizothoracine fishes in relation to the upheaval of the Xizang Plateau. *Science Press, Beijing*. 125–126 (1983).
7. He, D., Chen, Y., Chen, Y. & Chen, Z. Molecular phylogeny of the specialized schizothoracine fishes (Teleostei: Cyprinidae), with their implications for the uplift of the Qinghai-Tibetan plateau. *Chin Sci Bull* 49, 39–48 (2004).
8. Yonezawa, T., Hasegawa, M. & Zhong, Y. Polyphyletic origins of schizothoracine fish (Cyprinidae, Osteichthyes) and adaptive evolution in their mitochondrial genomes. *Genes & Genetic Systems* 89, 187–191 (2014).
9. Wang, X., Gan, X., Li, J., Chen, Y. & He, S. Cyprinid phylogeny revealed independent origins of the Tibetan Plateau endemic polyplid cyprinids and their diversifications related to the Neogene uplift of the plateau. *Science China Life Sciences* 59, 1–17 (2016).
10. Peer, Y. V. D., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10, 725–732 (2009).
11. Wittbrodt, J., Meyer, A. & Schartl, M. More genes in fish? *Bioessays* 20, 511–515 (1998).
12. Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic Timing of the Fish-Specific Genome Duplication Correlates with the Diversification of Teleost Fish. *Journal of Molecular Evolution* 59, 190–203 (2004).
13. Ravi, V. & Venkatesh, B. Rapidly evolving fish genomes and teleost diversity. *Current Opinion in Genetics & Development* 18, 544–550 (2008).
14. Yang, H. Y. et al. Status Quo of Fishery Resources in the Middle Reach of Brahmaputra River. *Journal of Hydroecology* 3, 120–126 (2010).
15. Qi, D. L., Zhao, Y., Tang, W. J. & Yang, C. Threatened fishes of the world: Oxygymnocypris stewartii (Lloyd 1908) (Cyprinidae). *Environmental Biology of Fishes* 86, 351 (2009).
16. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology* 35, 62–67 (2013).
17. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1, 18 (2012).
18. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* 12, 780–786 (2015).
19. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10, 563–569 (2013).
20. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One* 9, e112963 (2014).
21. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212 (2015).

**Figure 5.** Venn diagram of the number of genes with functional annotation using multiple public databases.
22. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067 (2007).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 27, 1754–1760 (2009).
24. Kent, W. J. BLAT—the BLAST-like alignment tool. Genome Research 12, 656–664 (2002).
25. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644–652 (2011).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35, W265–W268 (2007).
27. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21, i351–i358 (2005).
28. Tempel, S. Using and Understanding RepeatMasker. Methods Mol Biol 859, 29–51 (2012).
29. Bao, W., Koijma, K. J. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6, 11 (2015).
30. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573–580 (1999).
31. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGIPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Research 33, W686–W689 (2005).
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410 (1990).
33. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935 (2013).
34. Li, Y.-h. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology 32, 1045 (2014).
35. Hubbard, T. et al. The Ensembl genome database project. Nucleic Acids Research 30, 38–41 (2002).
36. Wang, Y. et al. The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation. Nat Genet 47, 625–631 (2015).
37. Gertz, E. M, Yu, Y. K, Agarwala, R, Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol 4, 41 (2006).
38. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. Genome Res. 14, 988–995 (2004).
39. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 9, R7 (2008).
40. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33, W465–W467 (2005).
41. Guigó, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. J Mol Biol 226, 141–157 (1992).
42. Trapnell, C., Käll, L., Kim, D. D. & Rinn, J. L. RNA–seq: an assessment of current estimation techniques. Genome Biology 13, R42 (2012).
43. Liu, H. P. et al. Assembled the genome; Q.-Q.L. and S.-J.X. assessed the assembly quality; N.W. and J.-H.G. carried out the repeat annotation and gene annotation. H.-P.L., W.-K.J., Q.-Y.L., and Z.-B.M. conceived the study. H.-P.L., R.-B.Y., X.-H.Y. and W.-K.J. designed the scientiﬁc objectives. Q.-Y.L. and Z.-B.M. managed the project; Y.-C.L., J.W., C.Z. and C.-W.Z. collected the samples and extracted the genomic DNA; N.W. and D.W. estimated the genome size and assembled the genome; Q.-Q.L. and S.-J.X. assessed the assembly quality; N.W. and J.H.G. carried out the repeat annotation and gene annotation. H.-P.L., S.-J.X., N.W., J.-H.G., and W.-K.J. wrote the manuscript. Also, all authors read, edited and approved the final manuscript.
44. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
45. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14, R36 (2013).
46. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protoc 7, 562–578 (2012).
47. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45, D158–D169 (2016).
48. Finn, R. D. et al. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Research 45, D319–D3199 (2016).
49. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research 42, D199–D205 (2014).
50. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).

Data Citations
1. NCBI Sequence Read Archive SRP156257 (2018).
2. NCBI Sequence Read Archive SRP158092 (2018).
3. GenBank QVTF000000000 (2018).
4. Wu, N. Figshare https://doi.org/10.6084/m9.figshare.7350365.v1 (2018).

Acknowledgements
This work was supported by the special finance of Tibet autonomous region (No. 2017CZZX003, 2017CZZX004 and XZNY-2018-C-040), the National Natural Science Foundation of China (No. 31560144 and 31602207), and the National Key Research and Development Program of China (No. 2016YFC1200500).

Author Contributions
H.-P.L., W.-K.J., Q.-Y.L., and Z.-B.M. conceived the study. H.-P.L., R.-B.Y., X.-H.Y. and W.-K.J. designed the scientific objectives. Q.-Y.L. and Z.-B.M. managed the project; Y.-C.L., J.W., C.Z. and C.-W.Z. collected the samples and extracted the genomic DNA; N.W. and D.W. estimated the genome size and assembled the genome; Q.-Q.L. and S.-J.X. assessed the assembly quality; N.W. and J.H.G. carried out the repeat annotation and gene annotation. H.-P.L., S.-J.X., N.W., J.-H.G., and W.-K.J. wrote the manuscript. Also, all authors read, edited and approved the final manuscript.

Additional Information
Competing interests: The authors declare no competing interests.

How to cite this article: Liu, H. P. et al. The sequence and de novo assembly of Oxygymnocypris stewartii genome. Sci. Data 6:190009 https://doi.org/10.1038/sdata.2019.9 (2019).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
