Genome survey of *Misgurnus anguillicaudatus* to identify genomic information, simple sequence repeat (SSR) markers, and mitochondrial genome

Guiyun Huang1,2 · Jianmeng Cao2 · Chen Chen3 · Miao Wang2 · Zhigang Liu2 · Fengying Gao2 · Mengmeng Yi2 · Gang Chen1 · Maixin Lu2

Received: 30 July 2021 / Accepted: 29 November 2021 / Published online: 22 January 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

**Background** The dojo loach *Misgurnus anguillicaudatus* is an important economic species in Asia because of its nutritional value and broad environmental adaptability. Despite its economic importance, genomic data for *M. anguillicaudatus* is currently unavailable.

**Methods and results** In the present study, we conducted a genome survey of *M. anguillicaudatus* using next-generation sequencing technology. Its genome size was estimated to be 1105.97 Mb by using K-mer analysis, and its heterozygosity ratio, repeat sequence content, GC content were 1.45%, 58.98%, and 38.03%, respectively. A total of 376,357 microsatellite motifs were identified, and mononucleotides, with a frequency of 42.57%, were the most frequently repeated motifs, followed by 40.83% dinucleotide, 7.49% trinucleotide, 8.09% tetranucleotide, and 0.91% pentanucleotide motifs. The AC/GT, AAT/ATT, and ACAG/CTGT repeats were the most abundant motifs among dinucleotide, trinucleotide, and tetranucleotide motifs, respectively. Besides, the complete mitochondrial genome was sequenced. Based on the Maximum Likelihood and Bayesian inference analyses, *M. anguillicaudatus* yingde in this study was the “introgressed” mitochondrial type. Seventy microsatellite loci were randomly selected from detected SSR loci to test polymorphic, of which, 20 microsatellite loci were assessed in 30 individuals from a wild population. The number of alleles (*Na*), observed heterozygosity (*Ho*), and expected heterozygosity (*He*) per locus ranged from 7 to 19, 0.400 to 0.933, and 0.752 to 0.938, respectively. All 20 loci were highly informative (*PIC > 0.700*). Eight loci deviated from Hardy–Weinberg equilibrium after Bonferroni correction (*P < 0.05*).

**Conclusions** This is the first report of genome survey sequencing in *M. anguillicaudatus*, genome information, mitochondrial genome, and microsatellite markers will be valuable for further studies on population genetic analysis, natural resource conservation, and molecular marker-assisted selective breeding.

**Keywords** *Misgurnus anguillicaudatus* · Genome characteristics · Microsatellite motifs · Mitochondrial genome · Microsatellite markers
Introduction

Next-generation, high-throughput sequencing (NGS) is an economical and powerful strategy for investigating genomic resources. Genome survey sequencing through NGS offers fundamental genomic information relating to genome size, heterozygosity ratio, and repeat contents. In recent years, genome survey sequencing has been extensively applied to accurately predict the whole genome characteristics of aquaculture species and provided favorable support for evolutionary biology and adaptation survey [1–5]. In addition to genome information, genome survey sequencing can also characterize a considerable amount of sequence data for molecular markers development, such as microsatellites and single nucleotide polymorphisms. Owing to the advantages, such as co-dominance, high polymorphism, easy automatic analysis, and random distribution within the genome, microsatellites (SSRs) are cost-effective and useful molecular tools in studies on genetic population and conservation, genetic breeding, and genetic map construction [6–8]. Recently, based on genome survey sequencing, increasing genome-wide SSR markers have been identified in aquaculture species, such as Charybdis feriatus [9], Sillago sihama [10], Marsupenaeus japonicus [2], and Hemibagrus wyckioides [11].

The dojo loach Misgurnus anguillicaudatus, which belongs to the Cobitidae family, is an important endemic freshwater species and is widely distributed in China, Japan, Korea, and other Southeast Asian countries. *M. anguillicaudatus* is an important scientific research object due to its unique physiological characteristics. Due to its natural variability in ploidy level, *M. anguillicaudatus* is regarded as an ideal animal in which to explore the biological origin, the evolutionary history of polyploidy, and unisexual reproduction [12, 13]. In addition, *M. anguillicaudatus* can also be recognized as a potential model species to explore the mechanism of accessory air-breathing function [14]. In addition to scientific value, it has become an important economic species in Asia because of its nutritional value and wide environmental adaptability. Despite its economic importance, the supply of dojo loaches for human consumption mostly depended on the exploitation of wild populations. With the increasing demand, over exploitation and habitat destruction have resulted in the decrease of natural populations in recent years. The lack of genetic and genomic resources in *M. anguillicaudatus* has limited aquaculture production. In this study, we aimed to (1) obtain genome information using genome survey sequencing, (2) describe microsatellite distribution patterns, (3) characterize the complete mitochondrial genome and construct phylogenetic analysis, (4) develop genome-wide SSR markers from genome sequencing data. These results will provide effective support for natural resource conservation, genetic diversity detection, and molecular marker-assisted selective breeding.

Material and methods

*M. anguillicaudatus* material and DNA extraction

All experimental procedures for *M. anguillicaudatus* handling were approved by the Animal Care and Ethics Committee of the Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences. *M. anguillicaudatus* wild populations were collected in May 2018 from Yingde County, Qingyuan City, Guangdong Province, China (24°35′1.31″N, 113°15′86.66″E). Caudal fins were collected and stored in 95% ethanol. One male adult *M. anguillicaudatus* was chosen as the material for genome survey sequencing. Thirty caudal fin samples were subjected to test SSR markers for polymorphisms.

The HiPure Mollusk DNA Mini Kit (Magen, China) was used to extract genomic DNA from caudal fin samples following the manufacturer’s instructions. The quality and the concentration of genomic DNA were determined by 1.2% agarose gel electrophoresis and the spectrophotometer (TianGen, China), respectively. Based on the research of Feng et al. [15], five EST-SSR markers (Supplementary Table S1) have been previously applied to identify the ploidy of *M. anguillicaudatus*. The *M. anguillicaudatus* used for genome survey sequencing was diploid (Fig. S1).

Whole-genome survey sequencing

Genomic DNA samples were randomly sheared into 350 bp fragments using an ultrasonicator (Covaris Inc.). The TruSeq DNA LT Sample Prep Kit (Illumina, USA) was used to construct libraries. DNA fragments of required lengths were recovered using electrophoresis before end repair, following which poly A-tails and sequencing adapters were added. The obtained fragments were purified and then performed PCR amplification for library preparation. Following the standard protocol, the Illumina Hiseq Xten platform was used to sequence the DNA library. The production of genomic data was carried out by the Qingdao OE Biotech Co., Ltd. (Qingdao, China).

K-mer analysis and genome assembly

After filtering out the adapters, poly (N) sequences, and low-quality reads, the high-quality clean data were used to perform K-mer analysis using GCE v1.0.0 software. Based on the 17-mer distribution, information on the peak depth and the number of 17-mers was obtained and used to estimate
genome size, repetitive sequences, and heterozygosity ratio. Then the genome size can be estimated by using the following formula: 
\[
\text{Genome size} = \frac{\text{K-mer num}}{\text{Peak depth}} \times K \times \text{Peak depth} \times \text{Q30 score} \times \text{GC content}.
\]
Additionally, based on K-mer analysis, the heterozygosity rate and proportion of repeat sequence were calculated according to the methods described by Liu et al. [16].

SOAPdenovo (v2.04) software [17] was used to assemble clean reads with K-mer = 41 by applying the de Bruijn graph structure. The clean reads were used to assemble contigs, and contigs were then used to assemble scaffolds depending on its paired-end information. The SOAPaligner v2.21 software [18] was applied to estimate the GC depth distribution.

**Microsatellite identification**

Microsatellite motifs with repeat motifs in the de novo draft genome were screened in the Perl script MIconSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa/). The search parameters were set as a minimum of 10, 6, 5, 5, and 5 repeats for detecting mono-, di-, tri-, tetra-, and hexanucleotide microsatellite motifs, respectively.

**Mitochondrial genome assembly and phylogenetic analysis**

The mitogenome sequence (NC_011209.1) downloaded from the NCBI dataset was as the reference to isolate the mitogenome sequence from the filtered clean data using NOVOPlasty v3.7.2. Based on genome sequencing data, only 845 bp mitochondrial genome sequence was identified. Finally, the complete mitochondrial genome of *M. anguillicaudatus* yingde was obtained by next-generation sequencing.

The mitogenome sequence of 23 Cobitidae species and 2 outgroup species (*Danio rerio* and *Cyprinus carpio*) were downloaded from the NCBI GenBank database and used for phylogenetic analyses. On the basis of 13 concatenated mitochondrial protein-coding genes (PCGs), we performed the phylogenetic analysis by the Maximum Likelihood (ML) and Bayesian inference (BI) methods using IQ-TREE and MrBayes 3.1.2 [19], respectively. PartitionFinder v1.1.1 [20] was used to select the best partitioning strategy and the optimal nucleotide substitution model for the data sets using the Bayesian information criterion (BIC). The result of the Bayesian information criterion revealed that GTR + I + G model are the best-fit model for the data sets. The ML phylogenetic tree was carried out using IQ-TREE software with 1000 replicates of bootstrap. The BI analysis was performed with MrBayes 3.1.2 software based on 10,000,000 generations.

**Characterization of polymorphic microsatellite loci**

Seventy SSR loci containing dinucleotide to pentanucleotide repeats were randomly selected from our genome survey data and primers were designed using Primer 5 software to amplify microsatellites. Firstly, 70 SSR loci were tested using ten individuals for polymorphism detection. Then primers of verified polymorphic loci were estimated in a wild population with 30 individuals. The microsatellite polymorphism analysis strategy was performed following the description in Schuelke’s article [21]. Briefly, four different universal adapter sequences labeled with the fluorochromes (FAM, VIC, NED, or PET dyes) were added to the 5’ end of each forward primer. Universal adapter sequences were shown in Table S2.

PCR was carried out in a 10 µL volume that contained 1.0 µL of buffer, 0.8 µL of dNTP mixture (2.5 mM), 0.6 µL of Mg²⁺, 0.05 µL of Ex-Taq DNA polymerase, 1 µL of DNA (100 ng/µL), 4.55 µL of ddH₂O, 1 µL of forward/reverse primer mixture (F: R = 2:50, 1.5 µM), and 1 µL of fluorescent-labeled primer (1.5 µM). PCR amplification was performed in a thermal lab cycler (SensoQuest, Germany) under the following conditions: initial denaturation at 94 °C for 5 min; 25 cycles at 94 °C for 30 s, appropriate annealing temperature for 1 min, and extension at 72 °C for 1 min; an additional 12 cycles at 94 °C for 30 s, annealing at 53 °C for 1 min, and extension at 72 °C for 1 min; and a final extension at 72 °C for 10 min. PCR products were subjected to capillary electrophoresis in a 3730 Genetic Analyzer (Applied Biosystems Inc., Foster City, CA, USA) using GeneScan™ 600 Liz® (ABI) as the size standard. The genotypes were determined using the software of GeneMarker V2.2.0.

**Data analysis**

Cervus 3.0.7 software was applied to calculate the number of alleles (*Na*), observed heterozygosity (*Ho*), expected heterozygosity (*He*), and the polymorphism information content (*PIC*) of polymorphic loci. Genepop 4.0 software was used to perform tests for deviations from Hardy–Weinberg equilibrium (HWE). Micro-Checker v.2.2.3 software was used for calculating the frequency of null alleles (*Fua*).

**Results**

**Genome size estimation and sequence assembly**

A total of 145.46 Gb of raw data were generated by sequencing the genome survey library (350 bp). After filtering and elimination, 135.22 Gb of high-quality clean sequences were generated with Q30 scores assigned to 93.17%, which was
approximately 122.3 × coverage. The effective rate and the error rate were 92.96% and 7%, respectively (Table S3).

A total of 135.22 Gb were used for K-mer analysis in *M. anguillicaudatus*. The 17-mer frequency distribution derived from the sequencing reads was plotted in Fig. 1A. The K-mer analyses indicated that the peak of the 17-mer distribution was 90, and the total K-mer count was 99,537,205,022, therefore, the genome size of *M. anguillicaudatus* was estimated to be 1105.97 Mb. The average GC content of the *M. anguillicaudatus* genome was 38.03% (Fig. 1B). The 17-mer analysis showed that the heterozygosity and repeat content of *M. anguillicaudatus* genome was 1.45% and 58.98%, respectively (Table S4).

A total of 4,120,965 contigs were assembled from the genome survey sequencing data with an N50 of 378 bp. Based on the contigs, the genome assembly contained 3,488,543 scaffolds, with an N50 of 551 bp (Table 1).

### Table 1 Statistics of the assembled contigs and scaffolds in *M. anguillicaudatus*

| Description                  | Contigs data   | Scaffold data  |
|------------------------------|----------------|---------------|
| Total length (bp)            | 1,110,202,680  | 1,176,790,755 |
| Number of sequences          | 4,120,965      | 3,488,543     |
| N50 length (bp)              | 378            | 551           |
| N60 length (bp)              | 289            | 389           |
| N70 length (bp)              | 200            | 279           |
| N80 length (bp)              | 147            | 172           |
| N90 length (bp)              | 118            | 128           |
| Number of N50                | 712,341        | 484,744       |
| Number of N60                | 1,048,358      | 739,119       |
| Number of N70                | 1,508,194      | 109,574       |
| Number of N80                | 2,170,190      | 164,087       |
| Number of N90                | 3,018,783      | 244,573       |

**Fig. 1** Results of genome survey using NGS. A K-mer (K = 17) analysis for estimate genome size and heterozygote frequency of *M. anguillicaudatus*. The x-axis is depth (X); the y-axis is the proportion that represents the frequency at that depth divided by the total frequency of all depths. B Guanine plus cytosine (GC) content and depth correlation analysis of *M. anguillicaudatus*. The x-axis represents the GC content and the y-axis is the sequencing depth. The distribution of sequence depth is on the right side, while the distribution of GC content is at the top.
Microsatellite motifs identification

Based on the assembled draft genome sequences, a total of 376,357 microsatellite motifs (SSRs) were identified from the genome. Consequently, the microsatellite distribution frequency of *M. anguillcaudatus* genome was calculated to be approximately 340.3 microsatellites per Mb. Among SSRs, mononucleotide repeats (42.57%) were the most frequent repeat motifs, followed by dinucleotides (40.83%), trinucleotides (7.49%), tetranucleotides (8.09%), pentanucleotides (0.91%), and hexanucleotides (0.10%) (Table 2). Short repetitive motifs were the dominant motifs, especially mono- and di-repeats, which accounted for 83.40%.

The number of penta- and hexa-nucleotide motif types was greater than 83, while that of tri- and tetra-nucleotides was less than 32 (Fig. 2A). Although the penta- and hexa-nucleotide SSR loci included more motif types than other motif types, the proportion of every motif type in penta- and hexa-nucleotide SSR loci was very few. The frequency distribution range of microsatellite repeats ranged from 5 to 10 repeats for trinucleotides, mostly from 5 to 9 repeats for tetranucleotides, from 5 to 8 repeats for pentanucleotides, and from 5 to 7 repeats for hexanucleotides (Fig. 2B). Of tri-

### Table 2 Microsatellites characteristics in the genome of *M. anguillcaudatus*

| SSR mining                  | Total   | Percent (%) |
|-----------------------------|---------|-------------|
| Total number of identified SSRs | 376,357 |             |
| Mono-nucleotides repeat     | 160,226 | 42.57       |
| Di-nucleotides repeat       | 153,665 | 40.83       |
| Tri-nucleotides repeat      | 28,191  | 7.49        |
| Tetra-nucleotides repeat    | 30,445  | 8.09        |
| Penta-nucleotides repeat    | 3438    | 0.91        |
| Hexa-nucleotides repeat     | 392     | 0.10        |

![Fig. 2](image-url) The motif types and frequency distribution of *M. anguillcaudatus* SSRs. **A** The motif types for mononucleotide, dinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide. **B** Frequency distribution by the number of repeats for trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide.
tetra-, penta-, and hexa-nucleotide repeats, the most common repeat numbers (5–6) accounted for more than 50% of the total repeats, and the repeat numbers 5–7 accounted for more than 78% of the total repeats. The motif types increased with increasing repeat length, and the frequency of repeats decreased because mutation rates were higher in longer repeats.

In dinucleotide SSR loci, AC/GT was the most abundant type of repeats, accounting for 46.52%, and CG/CG was the rare type (0.32%), as shown in Fig. 3A. Of the trinucleotide repeats, AAT/ATT was the most abundant, accounting for 64.01%, followed by ATC/ATG (10.27%) and AAC/GTT (9.53%), as shown in Fig. 3B. Of the tetranucleotide repeats, ACAG/CTGT was the most abundant motif (34.02%), followed by AGAT/ATCT (12.83%) and AAAT/ATTT (10.64%) (Table S5).

**Mitochondrial genome assembly and phylogenetic analysis**

We obtained a 16,464 bp long circular mitochondrial genome of *M. anguillicaudatus*, which was composed of 13 protein-coding genes (PCGs), 22 tRNA genes, 2 rRNA genes, and a non-coding hypervariable control region showing the typical teleosts mitogenomic arrangement (Fig. 4). The mitochondrial genome was AT-biased (58.05%) with the lowest content of G among the four bases and a positive A + T skew, which was identical to the characteristics of the mitochondrial genomes of vertebrates. Apart from protein-coding gene ND6, the other eight tRNAs (tRNA-Ala, tRNA-Asn, tRNA-Cys, tRNA-Tyr, tRNA-Ser, tRNA-Glu, tRNA-Pro, tRNA-Gln) were also located on the light strand, whereas all other genes were encoded on the heavy strand. In the complete mitogenome, eleven intergenic spacers were observed: of these, nine intergenic spacers were small intergenic spacers ranging from 1 to 13 bp with 32 bp total length. The largest two intergenic spacers were found in tRNA-Asn/tRNA-Cys (30 bp) occurring in the light strand and COX2/tRNA-Lys (27 bp) occurring in the heavy strand. Six overlapping regions were identified in a total of 26 bp in length. Gene overlaps were observed at 11 gene junctions (tRNA-Ile/tRNA-Gln, ATP8/ATP6, ATP6/COX3, Nd4L/ND4, ND5/ND6, and tRNA-Thr/tRNA-Pro) with overlaps ranging from 1 to 10 bp. Regarding start codons, COX1 started with GTG, whereas the other 12 protein-coding genes shared an identical initiation codon ATG. Concerning stop codons, COX3 and CYTB were inferred to terminate with an incomplete stop codon, with four of the protein-coding genes (ND2, ND3, ND4, and ND5) using TAG as a stop codon, and the other seven genes (ND1, COX1, COX2, ATP8, ATP6, ND4L, and ND6) sharing TAA, respectively (Table S6).

To better understand the taxonomic status of *M. anguillicaudatus* yingde, we performed the phylogenetic relationship of *M. anguillicaudatus* yingde with other Cobitidae species and two outgroup species as inferred by the entire mitogenome. The phylogenetic analyses trees constructed by ML and BI methods provided identical phylogenetic topologies (Fig. 5). All nodes in the ML and BI tree were strongly supported. With *Cyprinus carpio* and *Danio rerio* as outgroups, the species from Cobitinae, Nemacheilinae, and Botiinae were monophyletic groups, then form one clade. However, within the Cobitinae, *M. anguillicaudatus* was polyphyletic and split into two
divergent clades. One clade included *M. anguillicaudatus* (NC_011209.1) and *M. anguillicaudatus* (DQ026434.1), and this clade clustered with other genus *Misgurnus* species (*M. nikolskyi* and *M. mohoity*). While in the other clade, *M. anguillicaudatus yingde* obtained in this study, *M. anguillicaudatus* (MF579257.1), and *M. anguillicaudatus* (HM856629.1) were clustered together, then this clade was closely related to genus *Cobitis* species, such as *Cobitis lutheri*, *C. macrostigma*, *C. sinensis*, and *C. striata*.

### Polymorphism of microsatellite loci

Among 70 microsatellite loci, 20 loci were detected to show polymorphism in a wild *M. anguillicaudatus* population of 30 individuals (Table 3). A total of 237 alleles were amplified, and the number of alleles (*Na*) per locus ranged from 7 (MA-SSR-19, MA-SSR-48) to 19 (MA-SSR-17), with an average of 11.85 alleles per locus. The observed heterozygosity (*Ho*) ranged from 0.400 to 0.933, and the average *Ho* was 0.638. The *Ho* value of MA-SSR-17 (0.933) was highest among all
loci, indicating high heterozygosity of this locus, while the $Ho$ value of MA-SSR-70 (0.400) was lowest in the detected loci. The average expected heterozygosity ($He$) was 0.877, ranging from 0.752 to 0.938. The result of $He$ indicated a high degree of genetic variation for these individuals.

Based on the classification of Botstein [22], the average value of the polymorphic information content ($PIC$) was 0.839, ranging from 0.700 (MA-SSR-20) to 0.917 (MA-SSR-17), indicating that 20 loci were all highly informative ($PIC > 0.500$). These results indicated that there was a relatively high level of genetic diversity in these loci. Hardy–Weinberg equilibrium (HWE) analysis showed that except that eight loci (MA-SSR-09, MA-SSR-10, MA-SSR-11, MA-SSR-35, MA-SSR-38, MA-SSR-43, MA-SSR-51, and MA-SSR-58) showed significant deviations from the HWE ($P < 0.05$), the other twelve loci were in conformity to HWE. These HWE deviations may be resulted from natural selection, gene mutation, genetic drift, and the presence of null alleles.

Discussion

With the great advance of NGS technology, genome survey sequencing has been successfully applied for the prediction of genome size and genome structural information for many non-model species without prior knowledge. Ploidy is an important factor that directly determines the genome size. As the increase in ploidy and chromosome numbers, genome size generally increases. Therefore, for polyploid organisms, the ploidy must be detected before the whole genome sequencing. *M. anguillicaudatus* has complex ploidy composition in the natural environment [23–25]. A previous study in natural populations of *M. anguillicaudatus* indicated that diploid (2N = 50) individuals were most common among Chinese *M. anguillicaudatus* populations [12]. In this study, *M. anguillicaudatus* used for 2b-RAD sequencing was diploid, which was identified by microsatellite markers. According to the reported genome data, except for few fish species [26–28], the genome size of most fish was generally less than 1 Gb. The estimated genome size of diploid *M. anguillicaudatus* (1105.97 Mb) was larger than that in most fishes, such as *Oryzias latipes* (700.4 Mb) [29], *Acanthogobius ommaturus* (928.01 Mb) [30], and *Hemibagrus wyckioides* (728 Mb) [11], which was related to a higher number of repetitive sequences.

In addition to ploidy, many other factors including heterozygosity ratio, repeat content, whole genome duplication also affect the quality of genome assembly. For genome sequencing, the K-mer analysis combined with genome survey sequencing has been successfully applied for the prediction of genome size and genome structural information for many non-model species without prior knowledge. Ploidy is an important factor that directly determines the genome size. As the increase in ploidy and chromosome numbers, genome size generally increases. Therefore, for polyploid organisms, the ploidy must be detected before the whole genome sequencing. *M. anguillicaudatus* has complex ploidy composition in the natural environment [23–25]. A previous study in natural populations of *M. anguillicaudatus* indicated that diploid (2N = 50) individuals were most common among Chinese *M. anguillicaudatus* populations [12]. In this study, *M. anguillicaudatus* used for 2b-RAD sequencing was diploid, which was identified by microsatellite markers. According to the reported genome data, except for few fish species [26–28], the genome size of most fish was generally less than 1 Gb. The estimated genome size of diploid *M. anguillicaudatus* (1105.97 Mb) was larger than that in most fishes, such as *Oryzias latipes* (700.4 Mb) [29], *Acanthogobius ommaturus* (928.01 Mb) [30], and *Hemibagrus wyckioides* (728 Mb) [11], which was related to a higher number of repetitive sequences.

In addition to ploidy, many other factors including heterozygosity ratio, repeat content, whole genome duplication also affect the quality of genome assembly. For genome sequencing, the K-mer analysis combined with genome survey sequencing has been successfully applied for the prediction of genome size and genome structural information for many non-model species without prior knowledge. Ploidy is an important factor that directly determines the genome size. As the increase in ploidy and chromosome numbers, genome size generally increases. Therefore, for polyploid organisms, the ploidy must be detected before the whole genome sequencing. *M. anguillicaudatus* has complex ploidy composition in the natural environment [23–25]. A previous study in natural populations of *M. anguillicaudatus* indicated that diploid (2N = 50) individuals were most common among Chinese *M. anguillicaudatus* populations [12]. In this study, *M. anguillicaudatus* used for 2b-RAD sequencing was diploid, which was identified by microsatellite markers. According to the reported genome data, except for few fish species [26–28], the genome size of most fish was generally less than 1 Gb. The estimated genome size of diploid *M. anguillicaudatus* (1105.97 Mb) was larger than that in most fishes, such as *Oryzias latipes* (700.4 Mb) [29], *Acanthogobius ommaturus* (928.01 Mb) [30], and *Hemibagrus wyckioides* (728 Mb) [11], which was related to a higher number of repetitive sequences.

In addition to ploidy, many other factors including heterozygosity ratio, repeat content, whole genome duplication also affect the quality of genome assembly. For genome sequencing, the K-mer analysis combined with genome survey sequencing has been successfully applied for the prediction of genome size and genome structural information for many non-model species without prior knowledge. Ploidy is an important factor that directly determines the genome size. As the increase in ploidy and chromosome numbers, genome size generally increases. Therefore, for polyploid organisms, the ploidy must be detected before the whole genome sequencing. *M. anguillicaudatus* has complex ploidy composition in the natural environment [23–25]. A previous study in natural populations of *M. anguillicaudatus* indicated that diploid (2N = 50) individuals were most common among Chinese *M. anguillicaudatus* populations [12]. In this study, *M. anguillicaudatus* used for 2b-RAD sequencing was diploid, which was identified by microsatellite markers. According to the reported genome data, except for few fish species [26–28], the genome size of most fish was generally less than 1 Gb. The estimated genome size of diploid *M. anguillicaudatus* (1105.97 Mb) was larger than that in most fishes, such as *Oryzias latipes* (700.4 Mb) [29], *Acanthogobius ommaturus* (928.01 Mb) [30], and *Hemibagrus wyckioides* (728 Mb) [11], which was related to a higher number of repetitive sequences.

In addition to ploidy, many other factors including heterozygosity ratio, repeat content, whole genome duplication also affect the quality of genome assembly. For genome
assembly, if the heterozygosity rate exceeds 0.5%, it is difficult to assemble, and if it exceeds 1%, it is even more difficult [31]. The heterozygosity rate of *M. anguillicaudatus* was approximately 1.45%, indicating a complex genome with the higher heterozygosity rate. Accordingly, we supposed that the low quality of assembled genome sequences

| Primers   | Sequences 5′–3′ | Motifs | Allele size (bp) | Na   | Ho   | He   | PIC  | $P_{HWE}$ | $F$ (Null) |
|-----------|-----------------|--------|-----------------|------|------|------|------|----------|------------|
| MA-SSR-04 | F:GAGCTCCACATACACACACAGA R:ACACATTTAATACCCTTCTGCTT | (ATAG)24 | 248–292 | 11 | 0.500 | 0.881 | 0.839 | 0.070 | 0.218 |
| MA-SSR-09 | F:AGAACGAGAGGCTAGGACTATTG R:TTCCTACAGCATACACACCACACAGA | (GA)57 | 186–300 | 12 | 0.579 | 0.912 | 0.878 | 0.001** | 0.212 |
| MA-SSR-10 | F:AGGCTTTATCTTTGTTGAGCAGA R:GTTCCCATGCAAACACAGGCAAAGT | (CTAT)23 | 170–262 | 13 | 0.524 | 0.925 | 0.894 | 0.000*** | 0.2668 |
| MA-SSR-11 | F:TGAAGAGAGGGCCGAGAGACCTACT | (TAGA)29 | 256–308 | 10 | 0.571 | 0.847 | 0.814 | 0.000*** | 0.1717 |
| MA-SSR-17 | F:ACACAGGCTTCTTGGTAAATA R:TCCCTTTGGCAGAATGGACCTC | (GAAT)8 | 185–290 | 19 | 0.933 | 0.938 | 0.917 | 0.093 | 0.0055 |
| MA-SSR-19 | F:AAATTAACGACAGAAGGCAACT | (ATTCA)10 | 206–251 | 7 | 0.600 | 0.752 | 0.700 | 0.297 | 0.0681 |
| MA-SSR-31 | F:GGTGCTGGAATAAAGTATCGG R:CCTCCACCTTCTCTCTCCTCG | (AGACC)8 | 195–330 | 10 | 0.478 | 0.753 | 0.702 | 0.113 | 0.2016 |
| MA-SSR-34 | F:TGTAGACACGTACAAACTGAR R:TAGCCTCAATGCAACCCTTCA | (TATC)9 | 296–372 | 15 | 0.724 | 0.897 | 0.871 | 0.050 | 0.1015 |
| MA-SSR-35 | F:CAGCTCATATTGTCTTGCAGAC R:GTCGCGGTATCCCTTATGCTT | (GATA)13 | 218–290 | 15 | 0.680 | 0.920 | 0.894 | 0.000*** | 0.1401 |
| MA-SSR-38 | F:GGGTACATTATCCACTTGACGC R:TCTGGACACCCCTATCCTTCG | (GAAT)9 | 345–393 | 10 | 0.435 | 0.894 | 0.862 | 0.001** | 0.335 |
| MA-SSR-39 | F:TGCAAGACACATATGTAAGA R:GAGGAACGAACTTCTCTCTCA | (GATA)9 | 226–282 | 14 | 0.625 | 0.911 | 0.883 | 0.425 | 0.1775 |
| MA-SSR-40 | F:ATCACATGCTTCTTCTCTGCTC R:AAAGCACAACACCTTTAAAGACG | (TGTC)9 | 266–238 | 11 | 0.882 | 0.870 | 0.831 | 0.877 | -0.018 |
| MA-SSR-43 | F:CAATTTCTCAATGAGGAGCTGRC R:GGCCACAAATACACCTTCTCA | (TCTA)9 | 237–369 | 12 | 0.667 | 0.899 | 0.850 | 0.016* | 0.1417 |
| MA-SSR-48 | F:TGTCTTGGAGAGCTACCTCTC R:AGCTCTCGAATGTGATTTGAGG | (ATT)12 | 255–276 | 7 | 0.818 | 0.848 | 0.784 | 0.137 | 0.0109 |
| MA-SSR-51 | F:ACGGTAGGTGAAAACACGTG R:GTGGACCTGTTGTTGTTGAGG | (TAACA)8 | 210–290 | 11 | 0.529 | 0.904 | 0.865 | 0.000*** | 0.2468 |
| MA-SSR-58 | F:CAACAGACACACACACACACR:CTGGCTTCTTCTCTTCA | (ACAG)11 | 210–390 | 17 | 0.714 | 0.919 | 0.895 | 0.005** | 0.1186 |
| MA-SSR-67 | F:TTAGTGAGGGAGGAGGTGIGG R:TGCATGTGAGACCTTTTGAAR | (TATC)7 | 226–291 | 10 | 0.714 | 0.858 | 0.819 | 0.150 | 0.0845 |
| MA-SSR-68 | F:GCATGTGATGTTGTTGTCATC R:GTAGTGGTCAGGTGCTGTTT | (AAATA)7 | 195–375 | 8 | 0.583 | 0.859 | 0.801 | 0.323 | 0.1718 |
| MA-SSR-69 | F:GCACCTGTAATCTGAGG | (ACAG)10 | 223–287 | 16 | 0.800 | 0.909 | 0.881 | 0.705 | 0.0505 |
| MA-SSR-70 | F:TAGGGATAAACGACGGACCTGAGR:ACCTTTTGGCAGTCTTAC | (CTAT)14 | 152–280 | 8 | 0.400 | 0.846 | 0.795 | 0.089 | 0.3467 |

Mean 11.85 0.638 0.877 0.839

Primers for each locus including a forward primer (F), a reverse primer (R), Na number of alleles, Ho observed heterozygosity, He expected heterozygosity, PIC polymorphism information content, $P_{HWE}$ $P$ values for Hardy–Weinberg equilibrium corrected for multiple comparisons using the false discovery rate, $F$ (Null) frequency of null alleles

*Significantly deviated from Hardy–Weinberg equilibrium ($P < 0.05$), **Significantly deviated from Hardy–Weinberg equilibrium ($P < 0.01$), ***Significantly deviated from Hardy–Weinberg equilibrium ($P < 0.001$)
with short scaffolds might be caused by the high heterozygosity rate of *M. anguilllcaudatus*. Previous studies concluded sex determination type according to the differences in heterozygosity ratio between females and males [32, 33]. In the male heterogamety system, the heterozygosity ratio in males was larger than that in females, therefore, female individuals used for 2b-rad sequencing may be more reasonable to ensure high-quality genome sequences. Besides heterozygosity rate, the GC content is also one of the factors that directly affect sequence bias [34]. GC content that exceeds the normal range (>65% or <25%) may result in sequence bias on the Illumina sequencing platform, and eventually influence the quality of genome assembly seriously [35]. The average GC content of the *M. anguilllcaudatus* genome was approximately 38.03%, which was in the acceptable range and did not have an effect on genome sequence quality. This is the first report of genome survey in *M. anguilllcaudatus* and these genome survey data give a preliminary understanding of the genomic characteristics before large-scale genome sequencing. However, further genome study combined with PacBio and Hi-C techniques is needed to acquire more accurate high-quality genome information of *M. anguilllcaudatus*.

Identifying the microsatellite markers from genome survey data has the advantages of time- and cost-effective compared with traditional methods of developing microsatellites. It affords a fast method of isolating large numbers of microsatellite markers from genome-wide data. In the present study, except for mononucleotide repeats, the dinucleotide repeat motifs were the most frequent motifs and CG/CG dinucleotide microsatellites were the least abundant dinucleotide SSR loci, this result was congruent with other microsatellite repeats studies as described previously in *Danio rerio*, *Oreochromis niloticus*, and *Oreochromis latipes* [36]. This may be due to the methylation of cytosine into thymidine [37].

Apart from nuclear genome sequences, genome survey sequencing data also contain mitochondrial genome sequences [38]. In aquatic animals, Xu et al. [33] have assembled the mitochondrial genome of *Platycephalus* sp.1. on the basis of genome survey sequencing data. Although we failed to isolate the complete mitochondrial genome from genetic data because of the complexity of *M. anguilllcaudatus* genome. We still suggested that such a cost-effective method to generate mitochondrial genome should be extensively applied in future exploration of genome survey data. Previous studies on the mitochondrial genome have found that the *M. anguilllcaudatus* was divided into two clades and was considered as a polyphyletic group. One clade and other genus *Misgurnus* species got together into a monophyletic group [39], while the other clade was clustered with the genus *Cobitis* species [40–42]. With respect that *M. anguilllcaudatus* have two mtDNA lineages, Kitagawa et al. [43] stated mtDNA introgression hypothesis: the mitogenome of *M. anguilllcaudatus* individual in close association with the genus *Cobitis* species was recognized as “introgressed” mtDNA, while the mitogenome of other individual in close association with genus *Misgurnus* species was be considered as “non-introgressed” mtDNA. In terms of the hypothesis, we supposed that the *M. anguilllcaudatus* yingde is the “introgressed” mitochondrial type.

### Conclusion

In conclusion, we firstly assessed *M. anguilllcaudatus* genomic information including genome size, GC content, heterozygosity, and repeat sequence ratio, using genome survey sequencing and K-mer analysis. Our findings demonstrated that *M. anguilllcaudatus* genome was complex with high heterozygosity rate and repeated sequences. Genome-wide microsatellite motifs and 20 polymorphic microsatellite loci in the wild population were identified. Moreover, the mitochondrial genome was described and phylogenetic analysis was presented to explore its taxonomic status. All of this genetic information will be conducive for whole genome sequencing, genome size evolution, population genetic diversity, and genome-assisted breeding in *M. anguilllcaudatus*.

### Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1007/s11033-021-07037-x.

### Acknowledgements
This study was supported by China Agriculture Research System of MOF and MARA (CARS-46) and the National Key R&D Program of China (No. 2018YFD0901201).

### Author contributions
Conceived and designed the experiments: GC ML. Performed the experiments, analyzed the genetic data: GH. Contributed reagents/materials/analysis tools: JC CC MW ZL FG MY. Wrote the paper: GH ML. All authors read, revised, and approved the manuscript.

### Data availability
The data used to support the findings of this study are available from the article and its supplementary information.

### Declarations

#### Conflict of interest
The authors declare no conflict of interest.

#### Ethical approval
All experimental procedures for loach handling were approved by the Animal Care and Ethics Committee of the Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences.

#### Consent to participate
This study was conducted at Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences. The institution and all authors agreed to participate in this research.

#### Consent to publish
The Author confirms: that the work described has not been published before; that it is not under consideration for publica-
tion elsewhere; that its publication has been approved by all co-authors if any; that its publication has been approved explicitly by the responsible authorities at the institution where the work is carried out. The authors agree to publication in the journal Molecular Biology Reports.

References

1. Zhou W, Hu YY, Sui ZH et al (2013) Genome survey sequencing and genetic background characterization of Gracilariopsis lenta-neiformis (Rhodophyta) based on next-generation sequencing. PloS One 8(7):e69909
2. Lu X, Luan S, Kong J et al (2017) Genome-wide mining, characterization, and development of microsatellite markers in Marsupenaeus japonicus by genome survey sequencing. Chin J Oceanol Limn 35(1):203–214
3. Shi LL, Yi SK, Li YK (2018) Genome survey sequencing of red swamp crayfish Procambarus clarkii. Mol Biol Rep 45(5):799–806
4. Song H, Zhang YX, Yang MJ et al (2018) Genome survey on invasive veined rapa whelk (Rapana venosa) and development of microsatellite loci on large scale. J Genet 97:e79–e86
5. Li ZY, Tian CX, Huang Y et al (2019) A first insight into a draft genome of silver silago (Silago sihama) via genome survey sequencing. Animals 9(10):1–8
6. Romana-Eguia MRR, Ikedo M, Basiao ZU et al (2004) Genetic diversity in farmed Asian Nile and red hybrid tilapia stocks evaluated from microsatellite and mitochondrial DNA analysis. Aquaculture 236(1):131–150
7. O’Connell M, Wright JM (1997) Microsatellite DNA in fishes. Rev Fish Biol Fisher 7(3):331–363
8. Andersson L, Haley C, Ellegren H et al (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263(5154):1771–1774
9. Fang S, Wu R, Shi X et al (2020) Genome survey and identification of polymorphic microsatellites provide genomic information and molecular markers for the red crab Charybdis feriatus (Linnaeus, 1758) (Decapoda: Brachyura: Portunidae). J Crustacean Biol 40(1):76–81
10. Qiu B, Fang S, Ikhwanauddin M et al (2020) Genome survey and development of polymorphic microsatellite loci for Silago sihama based on Illumina sequencing technology. Mol Biol Rep 47(4):3011–3017
11. Zhou YL, Wu JJ, Wang ZW et al (2021) Microsatellite polymorphism and genetic differentiation of different populations screened from genome survey sequencing in red-tail catfish (Hemibagrus wyckioides). Aquac Rep 19:241–250
12. Zhong J, Yi S, Ma L et al (2001) 2019 Evolution and phylogeography analysis of diploid and polyploid Misgurnus anguillicaudatus populations across China. Proc Royal Soc B 286:20190076
13. Li YJ, Yin J, Wang JB et al (2008) A study on the distribution of polyploid loaches in China. Nippon Suisan Gakk 74(2):177–182
14. Huang SQ, Cao XJ, Tian XC et al (2016) High-Throughput sequencing identifies microRNAs from posterior intestine of loach (Misgurnus anguillicaudatus) and their response to intestinal air-breathing inhibition. PloS One 11(2):e0149123
15. Feng B, Yi SY, Zhang MM et al (2018) Development of novel EST-SSR markers for ploidy identification based on de novo transcriptome assembly for Misgurnus anguillicaudatus. PloS One 13(4):e0195829
16. Liu BH, Shi YJ, Yuan JY et al (2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quant Biol 35:62–67
17. Luo RB, Liu BH, Xie YL et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):1–6
18. Li RQ, Yu C, Li YR et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25(15):1966–1967
19. Hulsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755
20. Lanfear R, Calcott B, Ho SYW et al (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701
21. Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18(2):233–234
22. Botstein D, White RL, Skolnick M et al (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am J Hum Genet 32(3):314–331
23. Cui L, Abbas K, Yu Y et al (2013) First record of the natural occurrence of pentaploid loach, Misgurnus anguillicaudatus in Hubei Province, China. Folia Zool 62(1):14–18
24. Abbas K, Li MY, Wang WM et al (2009) First record of the natural occurrence of hexaploid loach Misgurnus anguillicaudatus in Hubei Province, China. J Fish Biol 75(2):435–441
25. Arai K (2003) Genetics of the loach, Misgurnus anguillicaudatus: recent progress and perspective. Folia Biol 51:107–117
26. Lien S, Koop BF, Sandve SR et al (2016) The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205
27. Xu F, Zhang XF, Wang XM et al (2014) Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nat Genet 46(11):1212–1219
28. Berthelot C, Brunet F, Chalopin D et al (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5(1):3657–3667
29. Kasahara M, Naruse K, Sasaki S et al (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447(7145):714–719
30. Chen BJ, Sun ZC, Lou FR et al (2020) Genomic characteristics and profile of microsatellite primers for Acanthogobius ommatus by genome survey sequencing. Bio Rep 40:BSR20201295
31. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770
32. Huang YQ, Jiang DN, Li M et al (2019) Genome survey of male and female spotted scat (Scatophagus argus). Animals 9(12):1–17
33. Xu SY, Zhang H, Gao TX (2020) Comprehensive whole genome survey analyses of male and female brown-spotted flathead fish Platypocephalus sp. 1. Genomics 112(6):4742–4748
34. Cheung M, Down TA, Latorre I et al (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Res 39(15):e103
35. Aird D, Ross MG, Chen WS et al (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12(2):1–14
36. Lei Y, Zhou Y, Price M et al (2021) Genome-wide characterization of microsatellite DNA in fishes: survey and analysis of their abundance and frequency in genome-specific regions. BMC Genom 22(4):421–421
37. Schorderet DF, Gartler SM (1992) Analysis of CpG suppression in methylated and nonmethylated species. PNAS 89(3):957–961
38. Ji D, Sun Z, Song N et al (2020) The complete mitochondrial genome of Jaydalia lineata (Perciformes, Apogonidae) obtained by next-generation sequencing. Mitochondrial DNA B 5(3):2507–2508
39. He SP, Gu X, Mayden RL et al (2008) Phylogenetic position of the enigmatic genus Psilorhynchus (Ostariophy: Cypriniformes):
evidence from the mitochondrial genome. Mol Phylogenet Evol 47(1):419–425

40. Cui Z, Zhou X, Yu Y et al (2014) Comparative analysis of mitochondrial genomes in distinct nuclear ploidy loach *Misgurnus anguillicaudatus* and its implications for polyploidy evolution. PloS One 9(3):e92033

41. Zeng LG, Wang JH, Sheng JQ et al (2012) Molecular characteristics of mitochondrial DNA and phylogenetic analysis of the loach (*Misgurnus anguillicaudatus*) from the Poyang Lake. Mitochondr DNA 23(3):187–200

42. Lee SY, Bang IC, Nam YK (2018) Complete mitochondrial genome of albino cyprinid loach, *Misgurnus anguillicaudatus* (Cypriniformes: Cobitidae). Conserv Genet Resour 10:507–510

43. Kitagawa T, Fujii Y, Koizumi N (2011) Origin of the two major distinct mtDNA clades of the Japanese population of the oriental weather loach *Misgurnus anguillicaudatus* (Teleostei: Cobitidae). Folia Zool 60(4):343–349

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