Journal of Genetics, Vol. 97, Suppl 1, December 2018, Online Resources, pp. e111–e116
https://doi.org/10.1007/s12041-018-0997-6

ONLINE RESOURCES

Isolation and characterization of microsatellite markers in a highland fish, *Pareuchiloglanis sinensis* (Siluriformes: Sisoridae) by next-generation sequencing

WEITAO CHEN[1,2] and SHUNPING HE[1]*

[1] *The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, Hubei, People's Republic of China*
[2] *Pearl River Fisheries Research Institute, CAFS, Guangzhou 510380, Guangdong, People's Republic of China*
*For correspondence. E-mail: clad@ihb.ac.cn.

Received 29 September 2017; revised 7 February 2018; accepted 6 March 2018; published online 30 October 2018

Abstract. *Pareuchiloglanis sinensis* (Siluriformes: Sisoridae) is an endemic and highland fish species which occurs only in some rivers of south-west China. In this study, the isolation and characterization of polymorphic microsatellite loci of this fish species by next-generation sequencing is described. A total of 9471 simple-sequence repeats (SSRs) were observed from RNA-seq data. One hundred and twenty primer pairs were chosen randomly and validated across 48 *P. sinensis* individuals collected from the Dadu river (a branch of the Yangtze river) of which 28 polymorphic microsatellite loci were detected. The number of alleles ranged from 2 to 14, with an average of seven alleles per locus. Twenty loci exhibited high polymorphism with the polymorphism information content (PIC) larger than 0.5. The mean observed and expected heterozygosity varied from 0.104 to 0.958 and 0.157 to 0.844, with an average of 0.583 and 0.613, respectively. The microsatellite markers characterized in the current study serve as a useful tool for the conversation genetic studies and population evaluation of *P. sinensis*.

Keywords. microsatellite markers; Dadu river; RNA-seq; *Pareuchiloglanis sinensis*.

Introduction

Pareuchiloglanis sinensis (Siluriformes: Sisoridae) is an endemic and highland fish species which only detected in some rivers of south-west China, i.e. the Jinsha river, the Dadu river and the Bailong river (figure 1) (Chu *et al.* 1999). In our study, 48 individuals were collected from the Dadu river, a branch of the Yangtze river. As of March 2014, a total of 26 dams were constructed, some are under construction or planned for the river, which poses a new threat to freshwater ecosystems and fish diversity in the Dadu river (https://www.wilsoncenter.org/publication/interactive-mapping-chinas-dam-rush). To facilitate a better understanding of the genetic diversity and population structure of *P. sinensis* for resource conservation, we isolated and characterized 28 polymorphic microsatellites of *P. sinensis* owing to the fact that microsatellites are the markers of choice for a variety of population genetic studies. Compared with the traditional methods of simple-sequence repeats (SSRs) marker development, next-generation sequencing is more cost efficient (Zheng *et al.* 2013; Liu *et al.* 2017). RNA-seq data were generated by Ma *et al.* (2016). In this study, to understand the population genetics of *P. sinensis*, we used unigenes assembled from RNA-seq for developing polymorphic SSRs with a Perl script, MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html).

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12041-018-0997-6) contains supplementary material, which is available to authorized users.

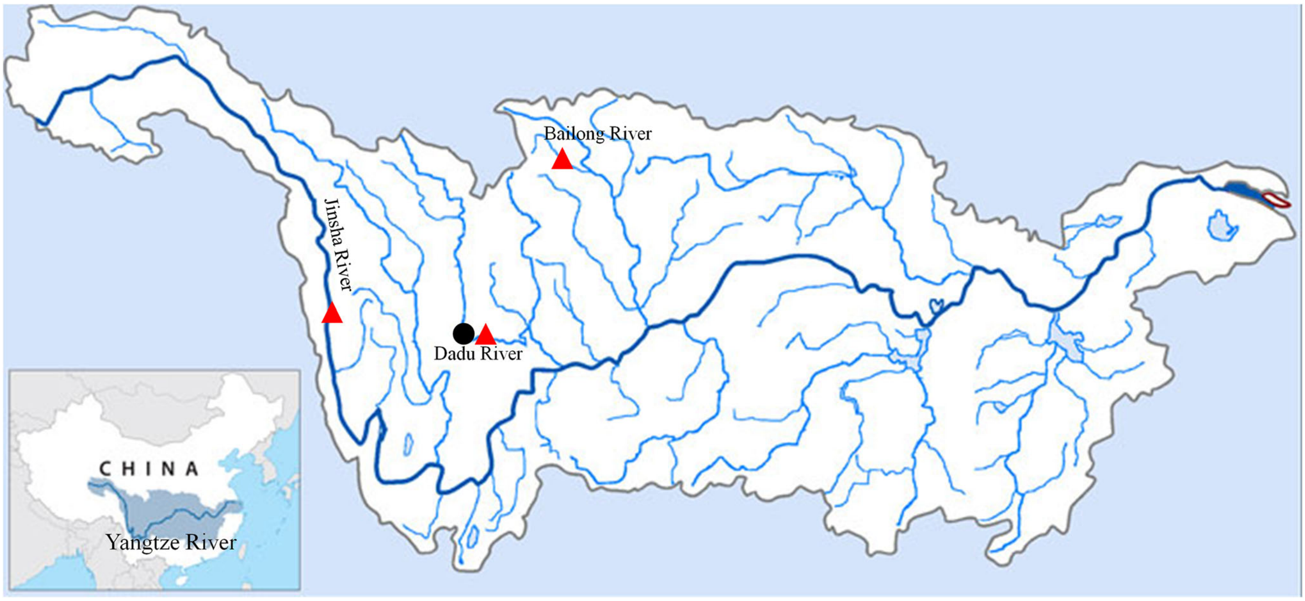

Figure 1. Distribution of natural species range and sampled locality for *P. sinensis*. Red triangle indicate natural species range of *P. sinensis*, and black dot represents sampled locality in the current study.

Materials and methods

Sample collection

In this study, methods involving fish were conducted in accordance with the Laboratory Animal Management Principles of China. Forty-eight individuals of *P. sinensis* were collected from the Dadu river.

RNA-seq data

RNA-seq data were generated by Ma *et al.* (2016). Fast QC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to control the quality of reads. We trimmed the adapter sequence and sites of lower quality reads (Phred score <20) with Cutadapt (Martin 2011). These cleaned reads were assembled using Trinity (Haas *et al.* 2013) software with default parameters. Contigs longer than 200 bp were retained for further analysis. CD–HIT–EST program (Li and Godzik 2006) with an identity threshold of 95% was used to remove low-coverage artifacts or redundancies. The unigenes were used for further microsatellite marker detection.

EST-SSR detection and primer development

Microsatellites within the unigene assembly were detected using a Perl script MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html). The SSR loci were considered to contain only two to six nucleotide motifs with a minimum of 6, 5, 5, 5 and 5 repeats, respectively. Mononucleotide repeats were excluded from the EST-SSR search as their polymorphism is often difficult to interpret (Lopez *et al.* 2015).

Table 1. Summary of SSRs identified in *P. sinensis* transcriptome unigenes.

Information	Number
Total number of sequences examined	47,989
Total size of examined sequences (bp)	37,172,544
Total number of identified SSRs	9471
Number of SSRs containing sequences	7832
Number of sequences containing more than one SSR	1354
Number of SSRs present in compound formation	492

The EST-SSR primers were designed using Primer 3.0 (Untergasser *et al.* 2012) under following criteria: (i) primers' length ranged from 18 to 25 bases (optimum: 20 bp); (ii) PCR product size ranged from 100 to 300 bp; (iii) melting temperature was between 58°C and 63°C (optimum: 60°C) and (4) a GC content of 40–60% (optimum: 50%).

DNA extraction, PCR conditions and amplification of SSRs

We dissected a small piece of white muscle tissue or fin from the right side of the body of each specimen. All of the tissue samples were preserved in 95% ethanol. Total genomic DNA was extracted from the muscle tissue or fin by performing a standard salt extraction.

The polymerase chain reaction (PCR) amplification was carried out in 30 μL reaction mixture with ~100 ng of

Table 2. Characterization of 28 transcriptome-derived microsatellites of *P. sinensis*. Primers redesigned from original sequence.

Locus	Primer sequence (5′–3′)	Repeat motif	Product size (bp)	T_m (°C)	N_A	H_O	H_E	P	PIC
BMK4	F: ACACACGCGTCTCTTCCTCT R: TGTGTGTCGGACCCTGAGACT	(GT)7	285–289	60	3	0.479	0.559	0.526	0.455
BMK9	F: ACATGCTTTTTACAAGCCCC R: GCCCCCAAAGAAGAAAGCTA	(ATC)6	152–164	60	6	0.875	0.721	0.598	0.659
BMK11	F: ACCGGAGTCTTTGGTCCTTT R: GAATTTGCCTCATTTCCCAA	(GTGA)5	272–316	60	9	0.729	0.745	**0.000**	0.698
BMK26	F: AGCATATCGGAAAGTGCCTG R: TTTCTTCTCCGCCGTTAAAA	(AC)7	249–253	60	3	0.104	0.157	**0.006**	0.148
BMK31	F: AGGCATCAAGCACATCAGTG R: AGGGAGATCTGGAGAGGGAG	(GT)8	236–264	60	8	0.458	0.511	**0.0207**	0.472
BMK43	F: ATGCAGAACACTCCCATTCC R: CATCAACGTGCTAATGTGCC	(GT)7	282–286	60	3	0.354	0.476	0.248	0.369
BMK46	F: ATGGTGAGTGCGCTACTGTG R: GGAAAGCAGCAAGCAGAAAA	(CA)8	104–114	60	6	0.729	0.713	0.644	0.654
BMK51	F: CAACAGCACGGTAGCTTCAA R: GTTGCTGAGCGGTCTCAGAT	(AT)8	118–134	60	7	0.708	0.734	**0.000**	0.684
BMK62	F: CACAGGTGTTCAGTCATCGG R: ATTAATCGTCCCCATTTCCC	(AC)7	262–300	60	11	0.688	0.774	0.066	0.738
BMK72	F: CATGTACAGGGGTTTGTGGG R: CAAATGCAAATGCAATCCAC	(TG)8	165–177	60	5	0.708	0.684	**0.000**	0.615
BMK74	F: CATGTCGATTCACAGTTCGG R: TACGCTCCTAACGTCTGCCT	(TTA)5	187–190	60	3	0.563	0.585	**0.001**	0.505
BMK81	F: CGAACGTGATCTCGAACTGA R: TCTGCAGGTCCATTTAGCAA	(TGT)6	262–271	60	3	0.271	0.274	1.000	0.248
BMK82	F: CGAAGAACTCTGATAGCGGG R: TGTAGAAGAAATGGGGGCTG	(CA)7	293–339	60	14	0.708	0.735	1.000	0.706
BMK89	F: CGATGAAGGTGTTGGTGATG R: GAAGGGATGACGCGAATTTA	(GAT)6	293–308	60	4	0.479	0.575	0.245	0.500

Table 2 (*contd*)

Locus	Primer sequence (5′–3′)	Repeat motif	Product size (bp)	T_m (°C)	N_A	H_O	H_E	P	PIC
BMK100	F: CTCAAAGAAACCTGAAGCCG R: TCTTGATGCGATACAAAGCG	(ACTA)5	240–256	60	4	0.188	0.192	0.080	0.179
BMK104	F: CTCCGCTCGTACACACTTCA R: TATGAACACACGCCCCAGTA	(CA)7	274–300	60	6	0.146	0.282	**0.000**	0.266
BMK118	F: CTGTATGGCTTGCGACAGAA R: GGAGGGTTGAAATGGGGTAT	(AG)8	255–287	60	11	0.750	0.797	**0.030**	0.762
BMK129	F: GACGAGAGCGAGAGAGAGGA R: GGAGAGAAAAGTTGGGGGAG	(AC)8	234–250	60	8	0.771	0.771		0.730
BMK132	F: GAGAAATGTGACACCTCGCA R: CTCTCTCTGTTCCGGTTTCG	(GT)8	274–300	60	13	0.958	0.844	**0.000**	0.816
BMK3	F: GAGCTGCTGGAAGAGTCACC R: TCGGAGCATTCCTTTCAGAT	(CT)8	287–317	60	13	0.833	0.884	0.944	0.862
BMK5	F: GAGCTTGGAGCAGAAAGCAG R: TCCCTCCTGAGCACTTGACT	(TTA)6	199–214	59	5	0.688	0.666	0.984	0.599
BMK6	F: GATGGCGTCTGAGTGTGAAT R: GTACATGCCGAACAACATGC	(AAG)6	112–124	59	4	0.500	0.451	0.808	0.394
BMK11	F: GCAAATGAAAAGCTGCGTAA R: CGGCACTGTAGGTCCTGTTT	(GT)9	174–194	59	8	0.771	0.658	**0.000**	0.591
BMK19	F: GCACTGCTACAACAGCGTTC R: TGTACCTGCCAACGTTCAAG	(TG)7	255–279	60	6	0.583	0.659	**0.000**	0.587
BMK21	F: GCATCGATCATTTCACATGG R: TTTGACAGCTAAGGCAGGAAA	$(TA)_8$	170–202	60	13	0.729	0.884	**0.000**	0.862
BMK28	F: GCTACACAGCCGAAGGAAAC R: GTCTTCTGTCTGGCTTTCGG	$(AC)_7$	235–275	60	11	0.521	0.710	**0.000**	0.673
BMK36	F: GCTCGTTCGCTTTGCTTTAC R: TTTCCACTTTCTCGCAATCC	$(TG)_8$	271–311	60	14	0.688	0.783	**0.030**	0.752
BMK38	F: GCTCTGTACAAGACCTCGCC R: TTCCCTGACTCGGATCAGTT	$(AAC)_5$	117–120	60	2	0.333	0.333	1.000	0.275

N_A, number of alleles; H_o, observed heterozygosity; H_E, expected heterozygosity; P, probability of deviation from Hardy–Weinberg equilibrium from heterozygote deficiency with significant values in bold.

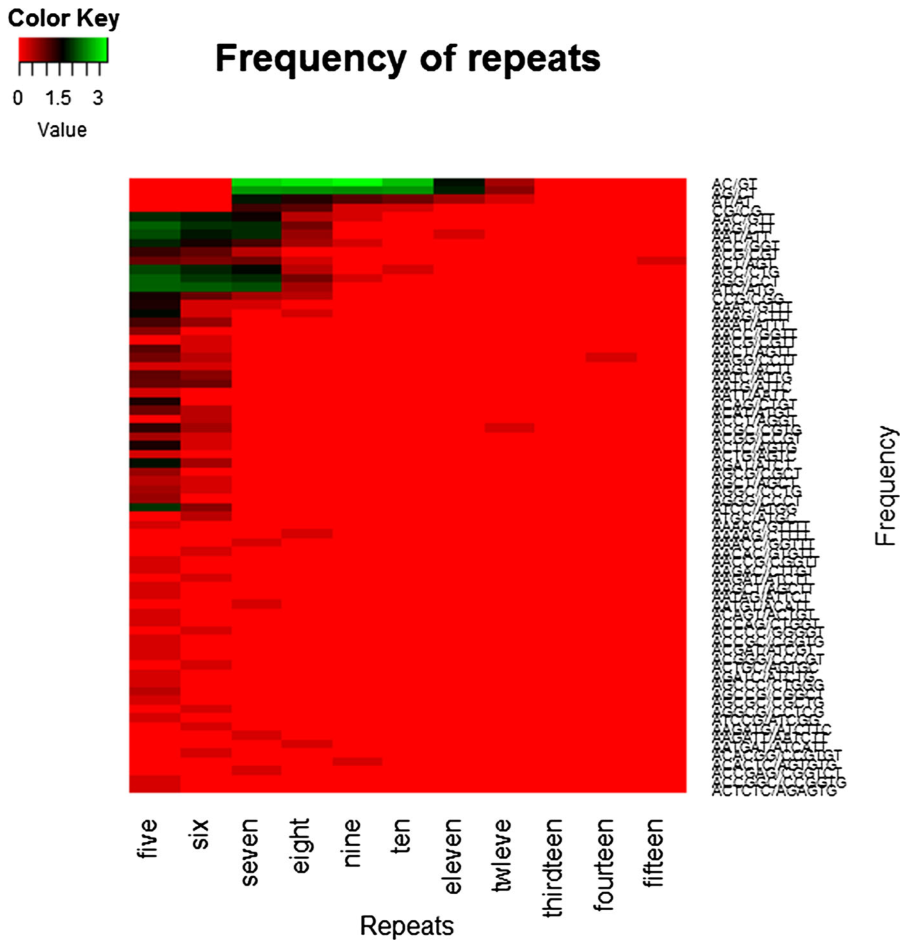

Figure 2. Heatmap of frequency of repeats identified by RNA-seq. Dinucleotide repeats were the most frequent (72.53%), followed by trinucleotide (22.56%) and tetranucleotide (4.56%) repeats. SSRs with nine tandem repeats (20.90%) were the most common.

template DNA, 1 μL of each primer (10 pmol), 3 μL of 10× reaction buffer, 1.5 μL of dNTPs (2.5 mM each) and 2.0 U of *Taq* DNA polymerase.

The PCR conditions for SSR included an initial denaturation step at 94°C for 5 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 60°C for 40 s and extension at 72°C for 30 s, followed by a final extension at 72°C for 10 min and storage at 4°C.

Amplification products were separated using 20% polyacrylamide gel. Some loci did not amplify in all samples although we adjusted the PCR conditions. These loci were excluded from further testing. Besides, only those loci which showed polymorphism were considered for genotyping analyses. Fluorescently labelled primers were further synthesized to ensure the accuracy of visualized lengths in polyacrylamide gel.

Genotyping

Forward primers (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/) were labelled with the FAM or HEX dye on the 5′-end. The PCR reaction conditions were the same as described above. The amplified products were detected on an ABI 3130xl Genetic Analyzer, and scored using GeneMapper software (Applied Biosystems, Foster City, USA).

Microsatellite data analysis

Important genetic parameters of polymorphic microsatellite loci such as polymorphism information content (PIC), the number of alleles (N_A), observed heterozygosity (H_O), expected heterozygosity (H_E) were calculated using POPGENE 1.32 (Quardokus 2000). Possible deviations from the Hardy–Weinberg equilibrium (HWE) were tested by Fisher's exact test with Bonferroni correction.

Results and discussion

In this study, 47,989 unigenes generated using RNA-seq data were used to detect potential microsatellite loci. A total of 7832 sequences were identified containing 9471

SSRs. A total of 1354 sequences contained more than one SSR (table 1). There were 70 motifs obtained, of which the most frequent was AC/GT (428, 54.65%), followed by AG/CT (406, 16.43%), ATC/ATG (138, 5.02%), AGG/CTT (123, 3.99%), AAG/CTT (101, 4.35%) and GTA/CAT (88, 3.79%) (table 1 in electronic supplementary material). Detailed analysis showed that dinucleotide repeats were the most frequent (72.53%), followed by trinucleotide (22.56%) and tetranucleotide (4.56%) repeats. SSRs with nine tandem repeats 1980 (20.90%) were the most common, followed by eight tandem repeats 1333 (14.07%) (figure 2).

To test the applicability and polymorphisms of SSR markers, 120 primer pairs were chosen randomly and validated across 48 *P. sinensis* individuals collected from the Dadu river (dot in figure 1). Of the 120 primer pairs only 86 (71.67%) were successfully amplified. Twenty-eight of the microsatellite loci showed polymorphism (table 2). Fluorescently labelled primers were further synthesized for these loci. The result showed that the number of alleles (N_A) for each locus ranged from 2 to 14 and the mean number of alleles per locus was 7. The observed heterozygosity (H_O) and expected heterozygosity (H_E) varied from 0.104 to 0.958 and from 0.157 to 0.844, with an average of 0.583 and 0.613, respectively (table 2). Twenty loci exhibited high polymorphism (PIC>0.5). Across all samples, 14 loci among 28 showed significant departures from the HWE (table 2).

P. sinensis is an endemic species with narrow distribution, which faced threat from human disturbance and habitat destruction. Thus, it is crucial that the current resources of *P. sinensis* be protected. Microsatellite markers developed in our study serve as a useful tool for the conversation genetic studies and population evaluation of *P. sinensis*.

Corresponding editor: Indrajit Nanda

Acknowledgements

This work was supported by the Key Fund and NSFC-Yunnan mutual funds of the National Natural Science Foundation of China (grant nos. 31130049 and U1036603).

References

Chu X., Zheng B. and Dai D. 1999 *Fauna Sinica, class Teleostei, Siluriformes* (in Chinese). Scientific Press, Beijing.

Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D., Bowden J. *et al.* 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512.

Li W. and Godzik A. 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659.

Liu H. G., Yang Z., Tang H. Y., Gong Y. and Wan L. 2017 Microsatellite development and characterization for *Saurogobio dabryi* Bleeker, 1871 in a Yangtze river-connected lake, China. *J. Genet.* **96**, e1–e4.

Lopez L., Barreiro R., Fischer M. and Koch M. A. 2015 Mining microsatellite markers from public expressed sequence tags databases for the study of threatened plants. *BMC Genomics* **16**, 781.

Ma X., Dai W., Kang J., Yang L. and He S. 2016 Comprehensive transcriptome analysis of six catfish species from an altitude gradient reveals adaptive evolution in Tibetan fishes. *G3-Genes Genomes, Genet.* **6**, 141–148.

Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12.

Quardokus E. 2000 PopGene. *Science* **288**, 458–458.

Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M. *et al.* 2012 Primer3 – new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115.

Zheng X., Pan C., Diao Y., You Y., Yang C. and Hu Z. 2013 Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). *BMC Genomics* **14**, 490.