IJA

Peng C, Chen G, Ye P, et al. Whole-Genome survey and microsatellite analysis of spotbanded scat *Selenotoca multifasciata*. *Israeli Journal of Aquaculture - Bamidgeh*. 2026;78(3):83-94. doi:10.46989/001c.165388

Original Research Articles

Whole-Genome survey and microsatellite analysis of spotbanded scat *Selenotoca multifasciata*

Chao Peng[1a], Gaojie Chen[1], Peirong Ye[1], Linwen Xie[1], Xianzhen Luo[1], Shuiqing Wu[2b], Sigang Fan[3]

[1] Hunan Provincial Key Laboratory for Molecular Immunity Technology of Aquatic Animal Diseases, College of life and environmental sciences, Hunan University of Arts and Science, Hunan Changde 415000, China, [2] Key Laboratory of Cultivation and High-value Utilization of Marine Organisms in Fujian Province，Fisheries Research Institute of Fujian, Xiamen 361000, China, [3] Key Laboratory of South China Sea Fishery Resources Exploitation & Utilization, Ministry of Agriculture and Rural Affairs, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, China

Keywords: Microsatellite (SSR), Genome survey, *Selenotoca multifasciata*

https://doi.org/10.46989/001c.165388

Israeli Journal of Aquaculture - Bamidgeh

Vol. 78, Issue 3, 2026

In this study, we conducted a comprehensive genome survey of *Selenotoca multifasciata* using Illumina short-read sequencing technology. A total of 51.63 Gb of high-quality clean sequencing data were generated, with Q20 and Q30 values reaching 98.77% and 96.64%, respectively. The 42.59 Gb of clean reads were assembled into 551,813 contigs (587.82 Mb) and 431,116 scaffolds (593.38 Mb). 17-mer frequency analysis estimated a genome size of 575.38 Mb, with 42.20% GC content, 0.43% heterozygosity, and 25.97% repeat ratio. 214,419 SSR loci were detected genome-wide, with dinucleotide repeats being the most prevalent type (80.49%) and AC/AG as the dominant motifs. Among 53 tested markers, 30 produced clear and stable bands, and 9 polymorphic loci were applied to assess the genetic diversity of a wild *S. multifasciata* population from Zhanjiang Bay. These results provide a valuable genomic basis for whole-genome sequencing and molecular marker development in *S. multifasciata* and related Scatophagidae species.

1. INTRODUCTION

Selenotoca multifasciata, commonly known as the spotbanded scat, belongs to the family Scatophagidae and is a highly valuable fish species with both edible and ornamental purposes.

It is naturally distributed in the coastal waters of Thailand, as well as the East China Sea and South China Sea of China.[1] Owing to its strong salinity tolerance, *S. multifasciata* can be cultured in both seawater and brackish water environments. As a marine omnivorous fish, it also exhibits broad tolerance to temperature and salinity.[2] Benefiting from its rapid growth rate and delicious flesh, *S. multifasciata* is widely cultured in net cages and ponds across China.[3] In recent years, significant progress has been made in various aspects of *S. multifasciata* research, including embryonic development,[3] larval rearing techniques,[2] polyculture systems,[4] mitochondrial genome characterization,[5] and the identification of sex-linked molecular markers.[1] With great aquaculture potential, *S. multifasciata* has not yet developed into a large-scale industry, leading to relatively low research attention. Meanwhile, whole-genome sequencing was prohibitively expensive, making it unaffordable for most research groups. Therefore, the genomic resources available for *S. multifasciata* remain extremely limited, severely hindering in-depth studies of its genetic diversity, population structure, and molecular breeding.

Next-generation sequencing (NGS) technology enables the rapid and efficient exploration of large-scale genomic resources.[6] These genomic resources are valuable for genetic breeding research, as they can provide functional genes associated with important traits, molecular markers, and targets for genome editing.[7]

While whole-genome sequencing and *de novo* assembly can provide the most comprehensive genomic information,[8] genome survey analysis using low-coverage NGS data has emerged as a practical and efficient approach to assess key genomic characteristics before undertaking large-scale whole-genome sequencing projects.[9-11] This approach allows researchers to estimate important genomic parameters such as genome size, heterozygosity level, GC content, and repeat sequence content, which are crucial for designing optimal sequencing strategies and selecting appropriate assembly algorithms. In recent years, genome survey analysis has been successfully applied to numerous marine fish species, including *Siganus oramin*,[11] *Scatophagus argus*,[12] and *Harpadon nehereus*.[13]

a Article Photo Cover: By RatioTile - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=164971273

b Corresponding author. Shuiqing Wu, e-mail: wushuiqing@163. com

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats, are tandemly repeated DNA sequences consisting of 2–6 base pairs (bp).[14] As co-dominant molecular markers with high polymorphism and genome-wide abundance, SSRs are widely used in marker-assisted selection breeding, species discrimination, genetic diversity studies, and population structure analysis.[14] For example, more than 299,574 and 299,893 SSRs (ranging from mono- to hexanucleotide repeats) have been detected in female and male *Scatophagus argus*, respectively.[12] Currently, there are no reports on SSR isolation in *S. multifasciata*.

Herein, a genome survey of *S. multifasciata* was conducted using NGS to elucidate critical genomic characteristics, including genome size, GC composition, and heterozygosity rate. Additionally, we conducted a comprehensive genome-wide identification and analysis of SSR loci and developed a set of polymorphic SSR markers. To our knowledge, this is the first report of SSR marker development in *S. multifasciata*. The results of this study will not only provide essential genomic resources for future whole-genome sequencing and assembly of *S. multifasciata* but also offer valuable molecular tools for conserving its genetic diversity and supporting sustainable aquaculture development.

2. MATERIALS AND METHODS

2.1. SAMPLE COLLECTION AND DNA EXTRACTION

A small wild specimen of *Selenotoca multifasciata* (body length = 3.61 cm) was purchased from the Huangsha Aquatic Products Market in Guangzhou in December 2024. Genomic DNA was extracted from ~30 mg of -80℃ preserved muscle tissue using the TIANamp Marine Animals DNA Kit (Tiangen, Beijing), and its quality and concentration were assessed by 1% agarose gel electrophoresis and Qubit Fluorometer (Thermo Fisher Scientific), respectively.

2.2. LIBRARY CONSTRUCTION AND ILLUMINA SEQUENCING

The qualified DNA sample was randomly sheared using an ultrasonic crusher (Covaris M220, Woburn, MA, USA). A paired-end (PE) library with an insert size of 350 base pairs (bp) was constructed from the randomly fragmented genomic DNA. Subsequently, the DNA library was sequenced on the Illumina NovaSeq sequencing platform (Illumina, San Diego, CA, USA) in PE 150-bp mode with a sequencing depth of 50×.

2.3. SEQUENCE ASSEMBLY AND ANALYSIS

Raw reads were processed with fastp 1.0[15] to remove adapter-contaminated, poly-N, and <75 bp reads, yielding clean reads for all subsequent analyses. Clean reads were assembled into contigs using SOAPdenovo v2.04 (k-mer=41; Luo et al.[16]), then scaffolded. For contamination check, 10,000 random clean reads were aligned to NCBI NT database via BLASTN (e-value<$1e^{-10}$, coverage>80%).

2.4. ESTIMATION OF GENOME SIZE, GC CONTENT, AND REPEAT RATIO

Genome size, heterozygosity and repeat content were estimated via 17-mer analysis using Jellyfish 2.1.4[17] and GenomeScope.[18]

2.5. MICROSATELLITE IDENTIFICATION AND ANALYSIS

Sequences with a length⩾300 bp were selected for SSR analysis. SSRs were screened using the MISA script with thresholds: ⩾6 repeats for dinucleotides, ⩾5 for tri- to hexanucleotides. The number, frequency, and motif types of SSRs were collected and statistically analyzed using WPS Office. Primers for SSR loci with⩾80 bp flanking sequences were designed using Primer3, with parameters: PCR product 100-330 bp, primer length 18-27 bp, GC content 40-60%, Tm 57-63℃.

2.6. SSR VALIDATION, GENOTYPING, AND DATA ANALYSES

Fifty-three SSR primer pairs were synthesized by Sangon Biotech (Shanghai, China) (Table S1). PCR was performed on a Bio-Rad T100 Thermal Cycler in 20-μL reactions containing 50 ng genomic DNA, 1 U Taq polymerase (Accurate, Changsha), 1×Mg^{2+} -containing buffer, 0.2 mM dNTPs, and 0.2 μM each primer. Cycling conditions: 94°C 5 min; 35 cycles of 94°C 30 s, 52–56°C (determined by the annealing temperature of the primer)

30 s, 72°C 30 s; 72°C 5 min. Amplicons were verified by 1% agarose gel electrophoresis. Validated forward primers were 5′-labeled with HEX/FAM, and fluorescent PCR products were analyzed by ABI 3730XL capillary electrophoresis with GeneMapper v6.0.

Thirty pairs of SSR primers yielded clear and stable amplification bands. Nine polymorphic SSR loci exhibiting clear fragment-length polymorphisms were empirically screened from eight individuals. We assessed the genetic diversity of 30 wild *S. multifasciata* individuals from Zhanjiang Bay using these nine SSR primer pairs. Genetic variation parameters, including the number of alleles (*Na*), number of effective alleles (*Ne*), observed heterozygosity (*Ho*), expected heterozygosity (*He*), and Shannon's Information Index (I), were calculated using Popgene 1.32.[19] The polymorphism information content (PIC) was determined using Cervus v3.0.7.[20]

3. RESULTS

3.1. GENOME SEQUENCING AND K-MER ANALYSIS

NGS generated a total of 51.63 Gb of raw data, corresponding to 172,112,276 raw reads (Table 1). After filtering and trimming low-quality sequences, more than 42.59 Gb of clean data (141,896,297 clean reads) were obtained (Table 1). The Q20 and Q30 values were 98.77% and 96.64%, re-

Table 1. Statistics of the genome survey sequencing data of *S. multifasciata*

Feature	Value
Raw reads	172,112,276
Raw Base (Gb)	51.63
Clean Reads	141,896,297
Clean Base (Gb)	42.59
Duplication rate (%)	15.66
Effective Rate (%)	82.44
Error Rate (%)	0.01
>Q20(%)	98.77
>Q30(%)	96.64
GC Content (%)	42.20

spectively, with a sequencing error rate of 0.01% (Table 1, Fig. 1a).

Genome survey analysis was performed using a 17-mer approach based on all clean reads. The K-mer depth distribution showed a main peak at a depth of 66, with a weak heterozygous peak observed between depths of 30 and 40 (Fig. 2). A total of 37,975,091,143 valid K-mers were generated (Table 2). Based on the K-mer distribution, the estimated genome size was 575.38 Mb, and the revised genome size was 567.48 Mb after filtering low-frequency erroneous K-mers (Table 2). The heterozygosity rate was calculated to be 0.43% (Table 2). In addition, the Num-Frequency curve exhibited a clear tailing pattern at depths greater than 100 (Fig. 2), and the repeat sequence ratio was estimated to be 25.97% (Table 2).

BLASTN analysis of randomly selected reads showed that the best hits were enriched in closely related fish species, including *Dicentrarchus labrax* (0.9%), *Haplochromis burtoni* (0.47%), *Takifugu rubripes* (0.29%), *Oreochromis niloticus* (0.28%), and *Xiphophorus maculatus* (0.16%).

3.2. GENOME ASSEMBLY AND SIZE

Details of contig and scaffold assembly for the draft genome are presented in Table 3, 551,813 contigs (587.82 Mb, N50=4,223 bp) were assembled into 431,116 scaffolds (593.38 Mb, N50=9,458 bp).

The coverage distribution of the assembled contigs was evaluated. The length distribution exhibited a prominent peak at 51×coverage, representing the effective sequencing depth of single-copy genomic regions (Fig. 3a). The number distribution also peaked at 51× coverage, confirming that most contigs were uniformly covered (Fig. 3b).

The GC content versus sequencing depth distribution was analyzed to assess data quality and genomic characteristics (Fig. 4). A prominent dense cluster was observed at GC% ≈ 40–50% and sequencing depth ≈ 30–60×. The unimodal GC histogram confirmed an average GC content of ~45% (Fig. 4). The sequencing depth histogram showed a main peak at ~50×. The GC-depth distribution was segregated into two layers: a low-depth layer below a high-depth layer (Fig. 4).

3.3. SSR IDENTIFICATION

A total of 214,419 SSR loci were identified from the assembled sequences (Table 4; Table S2). Dinucleotide repeats were the most abundant SSR type (172,576 loci, 80.49%), followed by trinucleotide (27,341 loci, 12.75%), tetranucleotide (10,753 loci, 5.01%), pentanucleotide (3,336 loci, 1.56%), and hexanucleotide repeats (413 loci, 0.19%) (Table 4). The overall microsatellite density was 377.84 loci/Mb, dominated by dinucleotide repeats (304.11 loci/Mb), followed by trinucleotide (48.18 loci/Mb) and tetranucleotide repeats (18.95 loci/Mb). Total SSR length was 1.15 Mb, representing 0.20% of the *S. multifasciata* draft genome (Table 4).

A total of 239 SSR motifs were detected in the assembled *S. multifasciata* genome, with 4, 10, 31, 88 and 106 types for di- to hexanucleotide repeats, respectively (Table S2). The repeat number of all motifs ranged from 5 to 19, with a general trend of decreasing copy number with increasing repeat unit length. Dinucleotide motifs were concentrated in 6–19 repeats, trinucleotide motifs in 5–13 repeats, and hexanucleotide motifs in 5–7 repeats (Table S2). The most abundant repeat number was 6, with 44,987 motifs (20.98%), followed by 7 repeats (29,892 motifs, 13.94%) and 8 repeats (21,939 motifs, 10.23%) (Table S2).

AC repeats were the most abundant SSR motif (136,135 loci, 239.89 loci/Mb), followed by AG repeats (28,147 loci, 49.60 loci/Mb) and AT repeats (8,192 loci, 14.44 loci/Mb) (Table 5), accounting for 78.88%, 16.31%, and 4.75% of dinucleotide repeats, respectively (Fig. 5). For trinucleotide repeats, AGG was the most abundant motif (7,262 loci, 12.80 loci/Mb), followed by AAT (6,416 loci, 11.31 loci/Mb) and AAG (3,290 loci, 5.80 loci/Mb) (Table 5), representing 26.56%, 23.47%, and 12.03% of trinucleotide repeats (Fig. 5). AGAT was the most abundant tetranucleotide motif (2,860 loci, 5.04 loci/Mb), accounting for 26.60% of tetranucleotide repeats (Fig. 5). AGAGG and ACAGAG were the dominant pentanucleotide and hexanucleotide motifs, respectively, although their proportions were relatively low (Fig. 5).

3.4. DEVELOPMENT OF GENOME-WIDE SSR MARKERS

Primers were designed for all identified SSR loci based on their flanking sequences, yielding 75,315 genome-wide SSR markers in *S. multifasciata* (Table 6). The frequency of SSR markers was 132.72 loci/Mb. Dinucleotide repeats were the most abundant type, with 57,333 markers (76.13% of the total SSR markers), followed by trinucleotide (13,731 markers, 18.23%) and tetranucleotide repeats (3,656 markers, 4.85%) (Table 6). In contrast, pentanucleotide (515 markers, 0.68%) and hexanucleotide repeats (75 markers, 0.10%) accounted for only a very small proportion of the total SSR markers.

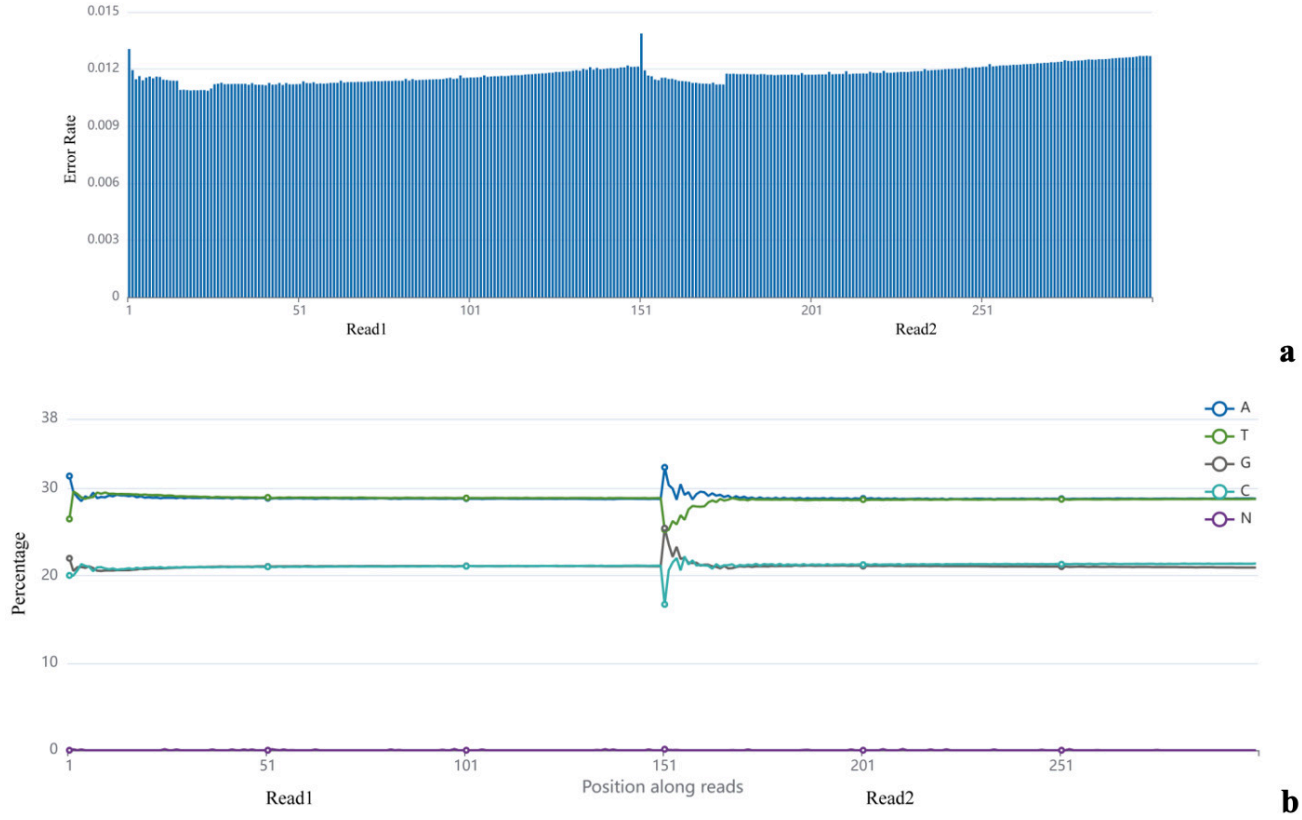

Fig. 1. Distribution of error rate (a) and GC content (b).

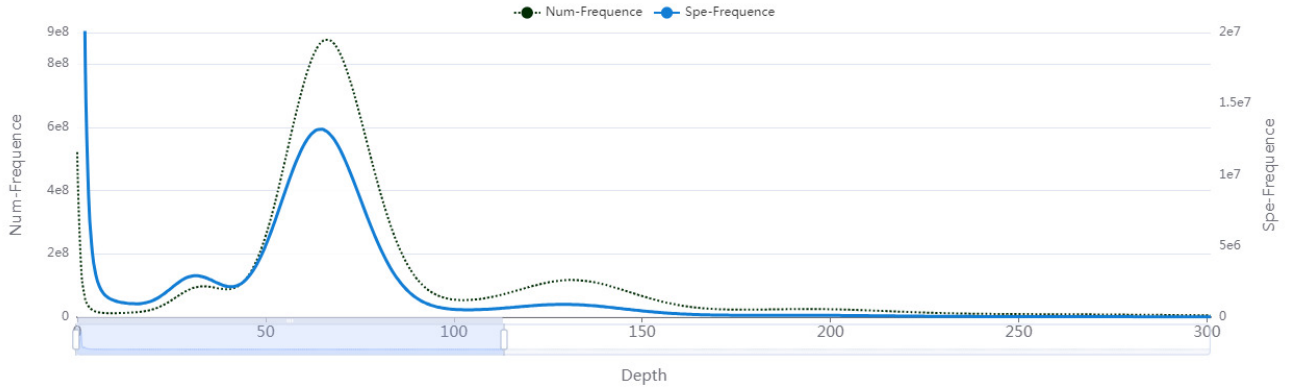

Fig. 2. K-mer (17-mer) distribution for estimating the genome size of *S. multifasciata*. The X-axis represents the K-mer depth, while the Y-axis denotes the number and types of K-mers at the corresponding depth. "Num-Frequency" and "Spe-Frequency" refer to the frequency distributions of K-mer counts and K-mer types, respectively.

3.5. VALIDATION AND POLYMORPHISM ANALYSIS OF SSR MARKERS

To evaluate the polymorphism of the developed SSR markers, 53 pairs of primers were synthesized (Table S1). Of the tested primer sets, 30 (56.60%) produced clear and reproducible bands after PCR amplification. Nine pairs of these primers were selected for capillary fluorescence electrophoresis to analyze 30 wild *S. multifasciata* individuals from Zhanjiang Bay (Table S1). Table 7 summarizes genetic diversity parameters for the 9 microsatellite loci. A total of 48 alleles were detected, with *Na*=2.0000–8.0000 (mean 5.3333), *Ne*=1.3846–3.5857 (mean 2.6190), *Ho*=0.20000–0.8667 (mean 0.4704), He=0.2825–0.7333 (mean 0.5827), *I*=0.4506–1.5274 (mean 1.1492), and PIC=0.2392–0.6866 (mean 0.5340). Most loci showed moderate to high polymorphism, suitable for population genetic analysis.

Table 2. Estimation of *S. multifasciata* genome based on 17-mer statistics

Feature	Value
K-mer	17
Depth	66
Number of K-mer	37,975,091,143
Genome size (Mb)	575.38
Revised Genome size (Mb)	567.48
Heterozygous rate (%)	0.43
Repeat rate (%)	25.97

Table 3. Statistics of genome assembly for *S. multifasciata*

	Contig	Scaffold
Total length (Mb)	587,82	593,38
Total number	551,813	431,116
Max length (bp)	203,621	375,119
N50 length (bp)	4,223	9,458
N90 length (bp)	397	494

4. DISCUSSION

4.1. GENOME SEQUENCING CHARACTERISTICS OF *S. MULTIFASCIATA*

S. multifasciata has become a valuable marine aquaculture species in the southeast coastal areas of China.[1] With breakthroughs in artificial breeding technology, large-scale cultivation of *S. multifasciata* has become feasible. However, research on this species remains limited, especially regarding the mining and utilization of its genetic resources. Genomic characteristics, including genome size, heterozygosity, GC content, and repeat ratio, can be efficiently analyzed using NGS and K-mer analysis without prior genomic information.[21] In recent years, this strategy has been successfully applied to genomic studies of several marine fish species, such as *Siganus oramin*,[11] *Gadus macrocephalus*,[22] and *Scatophagus argus*.[12] In the present study, the draft genome of *S. multifasciata* was sequenced and assembled, and key genomic parameters, including genome size, repeat content, heterozygosity, and GC content, were evaluated. The present genome survey was performed based on a single individual of *S. multifasciata*, which inevitably imposes several inherent limitations on genomic parameter estimation. All genomic characteristics calculated in this study, including genome size, heterozygosity, GC content, and repetitive sequence proportion, only reflect the genomic features of the sampled specimen rather than universal species-level or population-level genomic traits.[23,24] Owing to the lack of geographic population background and biological replicates, individual-specific genetic variants, rare mutations, and unique heterozygous site distributions may lead to biased parameter estimates. Furthermore, potential local population differentiation, germplasm introgression, and genetic drift cannot be ruled out, making it impossible to determine whether the observed genomic signatures reflect species-conserved characteristics or stochastic variation at the individual level.[23,24] Accordingly, these preliminary genomic estimates cannot serve as standardized baseline data for wild germplasm evaluation, population genetic analysis, or comparative genomic research of conspecific populations. Instead, they should only be regarded as fundamental reference information for future high-quality chromosome-level genome assembly and subsequent molecular marker development. To obtain more accurate and representative genomic parameters, future studies should include multiple individuals from different geographic populations to reduce individual genetic noise and validate population-level genomic characteristics.

Quality assessment of the raw sequencing data showed high Q20 (98.77%) and Q30 (96.64%) values (Table 1), indicating high data quality suitable for subsequent genome assembly and survey analysis. K-mer analysis has been widely used to estimate genome size in various animals.[11,25] In this study, the estimated genome size of *S. multifasciata* was 567.48 Mb (Table 2), which is larger than that of *Siganus oramin* (447.63 Mb; Huang et al.[11]) and *Sillago sihama* (508 Mb; Li et al.[26]), but smaller than that of *Scatophagus argus* (597.60 Mb; Huang et al.[12]) and *Gadus macrocephalus* (658.22 Mb; Ma et al.[22]).

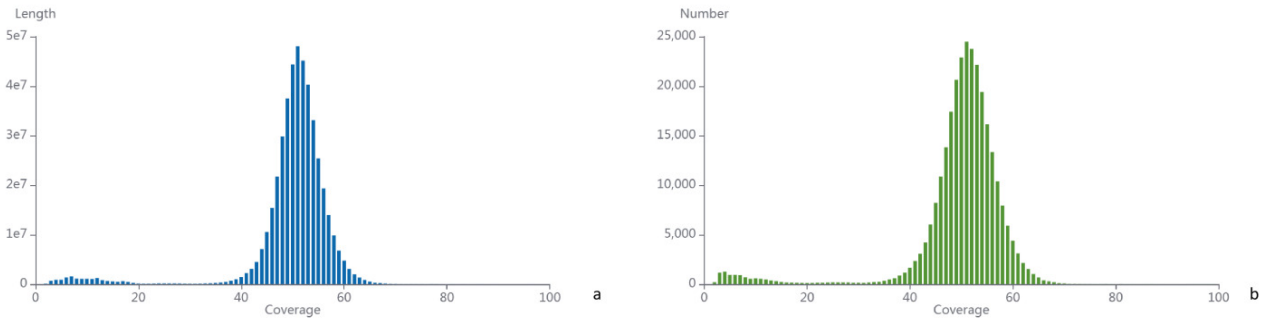

Fig. 3. Contig coverage depth and length distribution map (a), contig coverage depth and number distribution map (b).

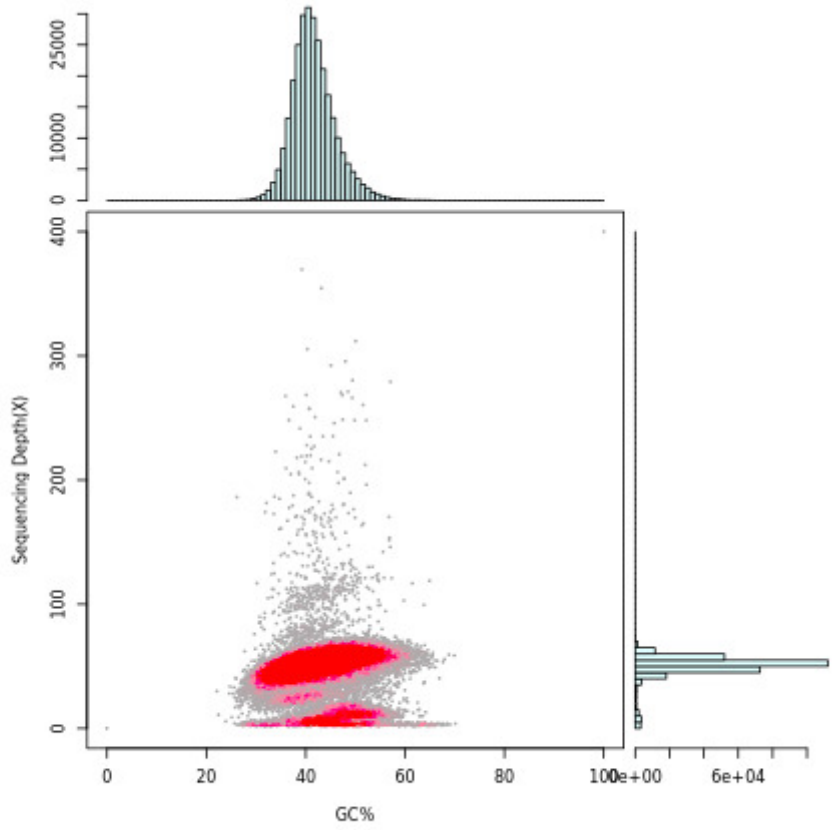

Fig. 4. GC content and depth-correlation analysis of *S. multifasciata*. The x-axis represents the GC content and the y-axis represents sequence depth. The distribution of sequence depth is on the right side. The distribution of GC content is at the top. The red and gray regions represent the dense of the scatter plot.

Table 4. SSRs in *S. multifasciata* genome

SSR motif	Number	Percentage (%)	Frequency (loci/Mb)	Length (bp)
Dinucleotide	172,576	80.49	304.11	209,017
Trinucleotide	27,341	12.75%	48.18	534,327
Tetranucleotide	10,753	5.01%	18.95	286,848
Pentanucleotide	3,336	1.56%	5.88	105,675
Hexanucleotide	413	0.19%	0.73	13,824
Total	214,419	100%	377.84	1,149,691

The genomic repeat ratio may affect genome size and chromosomal structure.[27] The Num frequency curve exhibited obvious tailing at sequencing depths greater than 100 (Fig. 2), whereas the length distribution showed no obvious tailing at high coverage (Fig. 3a). The repeat ratio of the *S. multifasciata* genome was 25.97%, representing a moderate level of repetitive sequences. This value is close to that of *Scatophagus argus* (27.06%; Huang et al.[12]), *Protosalanx hyalocranius* (24.43%; Liu et al.[28]), and *Siganus oramin* (28.01%; Huang et al.[11]), but lower than that of *Acanthogobius ommaturus* (38.31%; Chen et al.[29]) and *Pogonophryne albipinna* (39.03%; Jo et al.[25]).

Besides repetitive sequences, heterozygosity is another critical factor affecting genome assembly and analysis.[27] High heterozygosity often leads to fragmented assembly, redundant contigs, and overestimated genome size.[27] Assembly becomes particularly challenging when heterozygosity exceeds 0.5%.[17] In this study, a weak secondary peak was observed to the left of the main K-mer peak, and the estimated heterozygosity was 0.43%, indicating low to moderate heterozygosity. Moreover, the narrow and symmetric peak shape suggested uniform sequencing coverage and reliable assembly quality (Fig. 3b). Collectively, these results

Table 5. Top-10 repeat motifs found in the assembled genomic sequences of *S. multifasciata*

Repeat motifs	Number	Frequency (loci/Mb)
AC	136135	239.89
AG	28147	49.60
AT	8192	14.44
AGG	7262	12.80
AAT	6416	11.31
AAG	3290	5.80
AGAT	2860	5.04
AAC	2807	4.95
AGC	2557	4.51
ATC	2213	3.90

demonstrate that the *S. multifasciata* genome has relatively low heterozygosity, facilitating genome assembly.

Genomes with heterozygosity below 1% are generally considered easy to assemble.[17] Thus, the genome of *S. multifasciata* is amenable to high quality assembly. The GC content of most marine fish exceeds 40%, such as *Scatophagus argus* (41.78%; Huang et al.[12]), *Acanthopagrus latus* (42.07%; Zhu et al.[30]), and *Sillago sihama* (45.04%; Li et al.[26]). Consistent with these reports, the GC content of *S. multifasciata* was 42.20% (Table 1; Fig. 4), which falls within the favorable range for genome sequencing and assembly.[31]

In conclusion, the *S. multifasciata* genome exhibits a moderate size, low to moderate heterozygosity, and a moderate proportion of repetitive sequences, all of which are conducive to high quality genome assembly. Based on these findings, we propose that a high-quality chromosome-level genome of *S. multifasciata* can be constructed in future studies using PacBio long-read sequencing and Hi-C technology. Chromosome-level genome assembly enables precise characterization of chromosomal architecture, gene arrangement, repetitive sequence distribution, and large-scale structural variations, providing a fundamental resource for studying fish evolution, trait localization, stress adaptation mechanisms, and germplasm genetics.[32,33]

4.2. CHARACTERIZATION OF SSRS

SSR loci can be efficiently, accurately, and rapidly identified from genomic sequences. With decreasing costs of genome sequencing, large numbers of SSRs have been mined from genomic data in many species. To date, no studies on SSR development have been reported in *S. multifasciata*. In this study, a total of 214,419 SSR loci were identified from genome survey data. This number is lower than those reported in *A. ommaturus* (349,138), *G. macrocephalus* (789,860), and *H. nehereus* (760,890),[13,22,29] but higher than those in *Sillago sihama* (149,257; Li et al.[26]) and *Sebastiscus marmoratus* (191,592; Xu et al.[34]). Generally, the total number of SSRs is correlated with genome size.[22,26] In addition, the parameters used for SSR detection also influence the total number of identified loci. The density of microsatellites in the *S. multifasciata* genome was 377.84 loci/Mb, which is higher than that in *S. sihama* (293.81 loci/Mb; Li et al.[26]) and *A. ommaturus* (376.23 loci/Mb; Chen et al.[29]). This may be caused by the following reasons. First, intrinsic biological factors of the species play a decisive role, and replication slippage constitutes the primary driver of microsatellite expansion. Higher mismatch repair (MMR) activity leads to fewer accumulated SSRs and lower SSR density.[35] Second, technical artifacts from genome survey sequencing[36] introduce bias: insufficient sequencing depth and short read lengths fail to span long tandem microsatellites, resulting in extensive missed detection of long SSR loci.[37] Third, artificial SSR screening thresholds (minimum repeat copy number) directly alter statistical estimates of SSR density. The SSR filtering parameters used for S. sihama differ slightly from those used for *A. ommaturus* and *S. multifasciata*.[26,29]

Among all SSR types, dinucleotide repeats were the most abundant (80.49%, 304.11 loci/Mb) in the *S. multifasciata* genome, followed by trinucleotide (12.75%, 48.18 loci/Mb) and tetranucleotide repeats (5.01%, 18.95 loci/Mb). This pattern is consistent with previous studies in fish; for instance, dinucleotide repeats accounted for 74.42% in *Siganus oramin*[11] and 86.87% in *Pogonophryne albipinna*.[25] This phenomenon can be explained by the short repeat unit (2 bp) of dinucleotide motifs, which are more prone to DNA polymerase slippage during replication, leading to higher mutation rates and more frequent expansion. The AC motif was the most dominant dinucleotide repeat, followed by AG, which is consistent with observations in other fish species,[10,38] *Macaca mulatta*,[39] and mosquitoes.[40]

Owing to the limited length of scaffolds generated from genome survey sequencing, the maximum repeat number of SSRs in *S. multifasciata* was only 19, much lower than that detected in high quality assembled genomes. For example, SSR repeat numbers range from 5 to 447 in *Lateolabrax maculatus*,[38] 5 to 319 in *A. latus*,[10] and 5 to 259 in *Hypophthalmichthys molitrix*.[41] In addition, the total length of SSR regions in *S. multifasciata* was 1.15 Mb (Table 4), considerably lower than that in *L. maculatus* (5.12 Mb; Fan et al.[38]), *H. molitrix* (6.49 Mb; Wang et al.[41]), and *A. latus* (9.07 Mb; Peng et al.[10]). SSR analyses of these species were all based on complete genome assemblies. Therefore, obtaining a high quality reference genome of *S. multifasciata* will enable more comprehensive identification of SSR loci and deeper exploration of its genetic resources.

SSR markers were successfully developed by designing primers based on sequences flanking SSR loci.[14] In this study, 75,315 SSR markers were obtained, representing only 35.16% of all detected SSR loci. The low conversion rate was mainly attributed to the short flanking sequences of many SSR loci, which hindered the design of suitable primers. Subsequently, 53 markers were randomly selected and validated in 30 wild *S. multifasciata* individuals, and six highly polymorphic SSR markers (PIC > 0.5) were identified.[42]

In summary, this study performed a genome survey of *S. multifasciata* using K-mer analysis to estimate genome size, heterozygosity, GC content, and repeat ratio. Furthermore, we systematically characterized the distribution of SSRs and identified polymorphic SSR markers. These re-

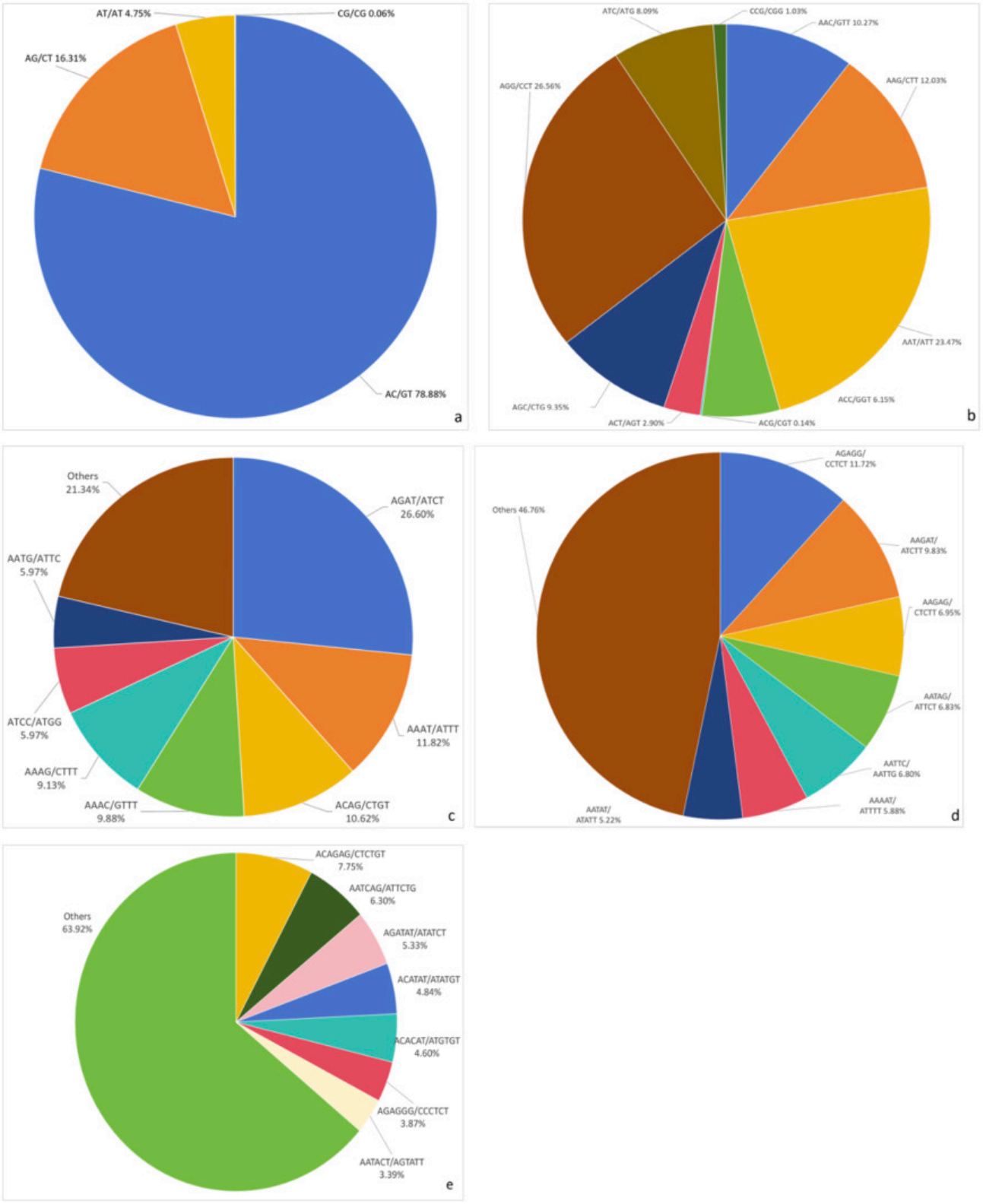

Fig. 5. Distribution of microsatellite motifs in the *S. multifasciata* genome. (a) Dinucleotide. (b) Trinucleotide. (c) Tetranucleotide. (d) Pentanucleotide. (e) Hexanucleotide.

sults provide a valuable basis for whole genome sequencing and assembly of *S. multifasciata*, as well as important genomic resources for its genetic breeding and improvement.

ACKNOWLEDGMENTS

This work is supported by the Open Program of Key Laboratory of Cultivation and High-value Utilization of Marine Organisms in Fujian Province (2025fjscq04); Hunan Natural Science Foundation Youth Project (2024JJ6331); the Excel-

Table 6. SSR markers in the *S. multifasciata* genome

SSR motif	Number of SSR markers	Percentage
Dinucleotide	57,338	76.13%
Trinucleotide	13,731	18.23%
Tetranucleotide	3,656	4.85%
Pentanucleotide	515	0.68%
Hexanucleotide	75	0.10%
Total	75,315	100%

Table 7. Parameters of genetic diversity of nine microsatellite loci

Locus	*Na*	*Ne*	*Ho*	*He*	*I*	PIC
Y15	7.0000	3.5857	0.3667	0.7333	1.5274	0.6866
Y23	5.0000	2.7607	0.6333	0.6486	1.2620	0.5993
Y24	4.0000	1.5762	0.4333	0.3718	0.7173	0.3398
Y40	6.0000	2.7692	0.5000	0.6497	1.3023	0.6014
Y14	2.0000	1.3846	0.2000	0.2825	0.4506	0.2392
Y25	5.0000	1.7982	0.2667	0.4514	0.8556	0.4000
Y38	5.0000	3.2550	0.4333	0.7045	1.3577	0.6461
Y44	8.0000	3.1746	0.5333	0.6966	1.4971	0.6506
Y41	6.0000	3.2668	0.8667	0.7056	1.3731	0.6429
Mean	5.3333	2.6190	0.4704	0.5827	1.1492	0.5340
St. Dev	1.7321	0.8212	0.1989	0.1682	0.3799	0.1630

Note: Na—observed number of alleles; Ne—effective number of alleles; Ho—observed heterozygosity; He—expected heterozygosity; I—Shannon's Information Index. PIC—polymorphism information content.

lent Young Scholars Project of Hunan Provincial Department of Education (25B0629); the Open Foundation of the Hunan Provincial Key Laboratory for Molecular Immunity Technology of Aquatic Animal Diseases (2025KF07).

AUTHORS' CONTRIBUTION

Conceptualization: Chao Peng (Lead). Formal Analysis: Chao Peng (Lead). Data curation: Chao Peng (Lead). Writing – original draft: Chao Peng (Lead). Methodology: Gaojie Chen (Equal), Peirong Ye (Equal), Linwen Xie (Equal), Xianzhen Luo (Equal). Writing – review & editing: Shuiqing Wu (Lead). Validation: Shuiqing Wu (Lead). Supervision: Shuiqing Wu (Lead). Project administration: Shuiqing Wu (Lead). Investigation: Sigang Fan (Lead).

COMPETING OF INTEREST – COPE

No competing interests were disclosed.

ETHICAL CONDUCT APPROVAL – IACUC

The study was conducted in accordance with the Declaration of Helsinki, and were approved by the Hunan University of Arts and Science Institutional Animal Care and Use Committee.

INFORMED CONSENT STATEMENT

All authors and institutions have confirmed this manuscript for publication.

DATA AVAILABILITY STATEMENT

All are available upon reasonable request.

SUPPLEMENTARY DATA

Table S1 Microsatellite markers used in this study.
Table S2 Summary of SSR motifs and repeats.

Submitted: August 20, 2026 CDT. Accepted: July 19, 2026 CDT.
Published: August 03, 2026 CDT.

REFERENCES

1. Jiang DN, Huang YQ, Zhang JM, et al. Establishment of the Y-linked Dmrt1Y as the candidate sex determination gene in spotbanded scat (*Selenotoca multifasciata*). *Aquac Rep*. 2022;23:101085. doi:10.1016/j.aqrep.2022.101085

2. Liu J, Ai T, Yang J, et al. Effects of Salinity on Growth, Digestive Enzyme Activity, and Antioxidant Capacity of Spotbanded Scat (*Selenotoca multifasciata*) Juveniles. *Fishes*. 2024;9:309. doi:10.3390/fishes9080309

3. Sun YQ, Liu JY, Zhuang P, et al. Effects of water temperature, salinity and pH on embryonic development of *Selenotoca multifasciata*. *South China Fish Sci*. 2021;17(6):122-129. doi:10.12131/20210109

4. Wang JJ, Luo DS, Liu JY. Nitrogen and phosphorus budgets of a polyculture system containing *Penaeus vannamei*, *Siganus guttatus*, and *Selenotoca multifasciata*. *J World Aquac Soc*. Published online 2023:1-17. doi:10.1111/jwas.12960

5. Liu Z, Mu X, Li H, et al. Complete mitochondrial genome of the striped scat *Selenotoca multifasciata* (Perciformes: Scatophagidae). *Mitochondrial DNA A DNA Mapp Seq Anal*. 2016;27(4):2691-2692. doi:10.3109/19401736.2015.1046125

6. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*. 2018;122(1):e59. doi:10.1002/cpmb.59

7. Woolley SA, Salavati M, Clark EL. Recent advances in the genomic resources for sheep. *Mamm Genome*. 2023;34:545-558. doi:10.1007/s00335-023-10018-z

8. Shi J, Tian Z, Lai J, et al. Plant pan-genomics and its applications. *Mol Plant*. 2023;16:168-186. doi:10.1016/j.molp.2022.12.009

9. Song N, Zhao XA, Cai CG, et al. Profile of the genomic characteristics and comparative studies of five Trichiuridae species by genome survey sequencing. *Front Mar Sci*. 2022;9:962307. doi:10.3389/fmars.2022.962307

10. Peng C, Luo C, Xiang G, et al. Genome-Wide Microsatellites in *Acanthopagrus latus*: Development, Distribution, Characterization, and Polymorphism. *Animals (Basel)*. 2024;14(24):3709. doi:10.3390/ani14243709

11. Huang XL, Li T, Yang YK, et al. Genome survey of *Siganus oramin*: Identification and development of genome-wide microsatellite markers. *Aquac Rep*. 2024;39:102520. doi:10.1016/j.aqrep.2024.102520

12. Huang Y, Jiang D, Li M, et al. Genome Survey of Male and Female Spotted Scat (*Scatophagus argus*). *Animals (Basel)*. 2019;9(12):1117. doi:10.3390/ani9121117

13. Yang T, Huang X, Ning Z, et al. Genome-Wide Survey Reveals the Microsatellite Characteristics and Phylogenetic Relationships of *Harpadon nehereus*. *Curr Issues Mol Biol*. 2021;43:1282-1292. doi:10.3390/cimb43030091

14. Wenne R. Microsatellites as Molecular Markers with Applications in Exploitation and Conservation of Aquatic Animal Populations. *Genes (Basel)*. 2023;14:808. doi:10.3390/genes14040808

15. Chen S. fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *Imeta*. 2025;4(5):e70078. doi:10.1002/imt2.70078

16. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1:18. doi:10.1186/2047-217X-1-18

17. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27:764-770. doi:10.1093/bioinformatics/btr011

18. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33:2202-2204. doi:10.1093/bioinformatics/btx153

19. Yeh FC, Boyle T. *POPGENE-1.32: A Free Program for the Analysis of Genetic Variation among and within Populations Using Co-Dominant and Dominant Markers*. University of Alberta; 2000.

20. Slate J, Marshall TC, Pemberton JM. A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Mol Ecol*. 2000;9:801-808. doi:10.1046/j.1365-294x.2000.00930.x

21. Hesse U. K-Mer-Based Genome Size Estimation in Theory and Practice. *Methods Mol Biol*. 2023;2672:79-113. doi:10.1007/978-1-0716-3226-0_4

22. Ma Y, Lou F, Yin X, et al. Whole-genome survey and phylogenetic analysis of Gadus macrocephalus. *Biosci Rep*. 2022;42:BSR20221037. doi:10.1042/BSR20221037

23. Romiguier J, Gayral P, Ballenghien M, et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*. 2014;515(7526):261-263. doi:10.1038/nature13685

24. Ellegren H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol*. 2014;29(1):51-63. doi:10.1016/j.tree.2013.09.008

25. Jo E, Cho YH, Lee SJ, et al. Genome survey and microsatellite motif identification of *Pogonophryne albipinna*. *Biosci Rep*. 2021;41:BSR20210824. doi:10.1042/BSR20210824

26. Li Z, Tian C, Huang Y, et al. A First Insight into a Draft Genome of Silver Sillago (*Sillago sihama*) via Genome Survey Sequencing. *Animals (Basel)*. 2019;9:756. doi:10.3390/ani9100756

27. Minio A, Cochetel N, Vondras AM, et al. Assembly of complete diploid-phased chromosomes from draft genome sequences. *G3 (Bethesda)*. 2022;12:jkac143. doi:10.1093/g3journal/jkac143

28. Liu K, Xu DP, Li J, et al. Whole genome sequencing of Chinese clearhead icefish, *Protosalanx hyalocranius*. *Gigascience*. 2017;6:1-6. doi:10.1093/gigascience/gix021

29. Chen B, Sun Z, Lou F, et al. Genomic characteristics and profile of microsatellite primers for *Acanthogobius ommaturus* by genome survey sequencing. *Biosci Rep*. 2020;40:BSR20201295. doi:10.1042/BSR20201295

30. Zhu KC, Zhang N, Liu BS, et al. A chromosome-level genome assembly of the yellowfin seabream (*Acanthopagrus latus*; Hottuyn, 1782) provides insights into its osmoregulation and sex reversal. *Genomics*. 2021;113:1617-1627. doi:10.1016/j.ygeno.2021.04.017

31. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 2011;12:R18. doi:10.1186/gb-2011-12-2-r18

32. Wu CS, Ma ZY, Zheng GD, et al. Chromosome-level genome assembly of grass carp (*Ctenopharyngodon idella*) provides insights into its genome evolution. *BMC Genomics*. 2022;23(1):271. doi:10.1186/s12864-022-08503-x

33. Liu J, Yin D, Ma F, et al. Telomere-to-Telomere Genome Assembly of Two Hemiculter Species Provide Insights into the Genomic and Morphometric Bases of Adaptation to Flow Velocity. *Biomolecules*. 2026;16(1):83. doi:10.3390/biom16010083

34. Xu SY, Song N, Xiao SJ, et al. Whole genome survey analysis and microsatellite motif identification of *Sebastiscus marmoratus*. *Biosci Rep*. 2020;40:BSR20192252. doi:10.1042/BSR20192252

35. Sinden RR. Origins of instability. *Nature*. 2001;411(6839):757-758. doi:10.1038/35081234

36. Xia Y, Luo W, Yuan S, et al. Microsatellite development from genome skimming and transcriptome sequencing: comparison of strategies and lessons from frog species. *BMC Genomics*. 2018;19(1):886. doi:10.1186/s12864-018-5329-y

37. Meglécz E, Nève G, Biffin E, et al. Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species. *PLoS One*. 2012;7(7):e40861. doi:10.1371/journal.pone.0040861

38. Fan SG, Huang H, Liu Y, et al. Genome-wide identification of microsatellite and development of polymorphic SSR markers for spotted sea bass (*Lateolabrax maculatus*). *Aquac Rep*. 2021;20:100677. doi:10.1016/j.aqrep.2021.100677

39. Xu Y, Hu Z, Wang C, et al. Characterization of perfect microsatellite based on genome-wide and chromosome level in Rhesus monkey (*Macaca mulatta*). *Gene*. 2016;592:269-275. doi:10.1016/j.gene.2016.05.046

40. Wang XT, Zhang YJ, Qiao L, et al. Comparative analyses of simple sequence repeats (SSRs) in 23 mosquito species genomes: identification, characterization and distribution (Diptera: Culicidae). *Insect Sci*. 2019;26:607-619. doi:10.1111/1744-7917.12647

41. Wang Y, Sha H, Li X, et al. Microsatellite Characteristics of Silver Carp (*Hypophthalmichthys molitrix*) Genome and Genetic Diversity Analysis in Four Cultured Populations. *Genes*. 2022;13:1267. doi:10.3390/genes13071267

42. Botstein D, White RL, Skolnick M, et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32(3):314-331.

SUPPLEMENTARY MATERIALS

Supplemental Table S1 & Table S2

Table S1 Microsatellite markers used in this study.
Table S2 Summary of SSR motifs and repeats.

Download: https://ija.scholasticahq.com/article/165388-whole-genome-survey-and-microsatellite-analysis-of-spotbanded-scat-_selenotoca-multifasciata_/attachment/354678.xlsx