Characterization of novel genotyping-by-sequencing (GBS)-based simple sequence repeats (SSRs) and their application for population genomics of *Capoeta aculeata* (Valenciennes, 1844)

Habibollah Gandomkar1 · Seyed Pezhman Hosseini Shekarabi1 · Hossein Ali Abdolhay2 · Sajad Nazari3 · Mehdi Shamsaie Mehrghan1

Received: 11 May 2021 / Accepted: 16 August 2021 / Published online: 22 August 2021 © The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

**Background** The species *Capoeta aculeata* (Valenciennes, 1844) is one of the most important freshwater species endemic to Iran. However, the investigation of the population genetic structure of this species is limited by the low number of molecular markers currently described.

**Methods and results** In this study, we implemented next generation sequencing technology to identify polymorphic microsatellite markers and investigate the population genetic structure of *C. aculeata* sampled from three geographical sites in Iran. We characterized and developed 36 novel polymorphic microsatellite markers and these loci were examined in 120 individuals from three populations occurring in the Zagros basin. The average number of alleles per locus varied from 1.7 to 16 (average = 7.89). The results showed that, the polymorphism information content (PIC) of these simple sequence repeat (SSR) loci varied from 0.254 to 0.888. The observed heterozygosity (*H*O) per locus ranged from 0.170 to 0.881, while the expected heterozygosity (*H*E) per locus was from 0.170 to 0.881. Among these SSR loci, 20 loci deviated significantly from the Hardy–Weinberg equilibrium after Bonferroni correction (*p* < 0.05).

**Conclusions** These microsatellite markers could provide a valuable tool for future population and conservation genetics studies of *C. aculeata* and other closely related species.

**Keywords** *Capoeta aculeata* · Microsatellite markers · Next generation sequencing (NGS) · Genetic structure

Introduction

The genus *Capoeta* is a herbivorous cyprinid genus and highly diversified with 14 species and widely distributed in water bodies of Western Iran. Kohgiluyeh and Boyer-Ahmad Province in the southwest part of Iran is a region with high number of endemism in some freshwater fish species including *Capoeta aculeata* (Valenciennes, 1844) [1]. However, populations of this species have experienced significant declines over the last 20 years, due primarily to habitat loss and fragmentation caused by water storage and land clearing. The previous phylogenetic and phylogeographic studies found that populations of *C. aculeata* are different from the *Capoeta* species [2–5].

Simple sequence repeat (SSR), also known as microsatellite markers are highly variable DNA sequences, distributed throughout eukaryotic genomes [6, 7]. Given its co-dominant inheritance, bi-parental mode of inheritance, polymerase chain reaction (PCR)-based, high amplification success rate and levels of polymorphism, microsatellite markers have been used in a wide range of applications in population genetics, molecular ecology, conservation genetics and quantitative trait loci (QTL) identification studies [8–10]. Nevertheless, the conventional methods for developing SSR markers are believed to be the key element which limits the application of microsatellite markers due to they...
are expensive, low efficiency, laborious and time-consuming [11, 12]. In fact, SSR markers are tremendously effective tools in population genetic studies because they might reveal the distinct population segments (DPSs) even in fine-scale genetic structure studies. Until now, microsatellite loci have not been reported in this species [13].

The major advantage of Next Generation Sequencing (NGS) technique is their ability to produce large amounts of sequence data and it can provide wide genetic information of simple sequence repeat (SSR) or microsatellite loci [14, 15]. Restriction-site associated DNA sequencing (RAD-seq) is a powerful tool to develop microsatellite and single nucleotide polymorphism (SNP) markers, which uses Illumina next-generation sequencing (NGS) platforms to discover informative volume of independent markers in a lot of specimens at the same time [16–18].

The reads created by RAD-seq can be grouped according to the enzyme recognition sequence, so this procedure could develop the accuracy and precision of contigs assembly, and consistently expand the success rate of polymorphic microsatellite markers development [19, 20]. In comparison with other techniques, restriction site-associated DNA sequencing (RAD-seq) gradually becomes the most promising method and can give genome-wide genotypic and sequence data adjacent to single restriction enzyme cut sites and generate reduced representation genomic libraries [21, 22]. Recently, the characterization and development of RAD-seq have been used to report microsatellite markers in many cyprinid species, such as Ctenopharyngodon Idella [23], Schizothorax prenanti [24] and Xenocypris argentea [25]. However, the genetic background of C. aculeata is limited and no microsatellite locus has been characterized in this species. In this study, RAD-seq was used to isolate and develop the first resource of microsatellite loci for C. aculeata from Iran.

Materials and methods

In this study, fin clips of 120 individuals were collected from three wild C. aculeata populations inhabiting the Beshar, Khersan, Maroun rivers in Kohgiluyeh and Boyer-Ahmad Province in Iran (Fig. 1, Table 1). All procedures were approved by the Committee on Animal Care and Use, and Committee on the Ethics of Animal Experiments of Agricultural Research, Education and Extension Organization (AREEEO) and Islamic Azad University. The modified method [26] was used to extract genomic DNA using the commercial DN-easy blood and tissue kit (Qiagen™, Germany) following the manufacturer’s instructions. The quality of DNA was assessed using a Nanodrop spectrophotometer (Nanodrop ND1000) and 1% agarose-gel electrophoresis, after which DNA concentration of each sample was diluted to 100 ng in a 100 μL volume and stored at –20 °C [27].

Screening of microsatellite loci by PCR

Total genomic DNA was isolated and purified from fin clips (about 30–50 mg) of each individual using the commercial DN-easy blood and tissue kit (Qiagen™, Germany) following the manufacturer’s instructions. Then, the final DNA concentration of each sample was standardized to 100 ng in a 10 μL volume [26]. The quality and concentrations of DNA samples were checked on 1.0% agarose gels and Nanodrop 1000 spectrophotometer, respectively.

RAD library construction

RAD library preparation was carried out following the protocol of [28]. The RAD sequencing of ten samples per each population was performed to develop microsatellite markers. Briefly, genomic DNA from each individual was digested for 2 h at 37 °C in a 40 μL reaction at specific sites with EcoRI-HF restriction enzymes according to the RAD protocol [29]. Adapter P1 containing individual-specific nucleotide barcodes was ligated to the digested product. Ligation reaction was performed in a 40 μL volume along with 5 μL of ligation buffer, 0.85 μL of P1 adapter, and 300 units of ligase for 2 h at 25 °C. DNA was subsequently ligated to a second adapter (P2), which contains T overhangs. The ligated materials were then pooled, purified, eluted and subjected to PCR enrichment. Total genomic DNA samples were sheared to fragment DNA to an average size of 350 bp and a pool of adapter-ligated molecules with random, variable ends was generated. DNA fragments with lengths of 250–500 bp were isolated using the Qiagen Min Elute Gel Extraction kit (Qiagen™, Germany). The libraries constructed were sequenced on the Illumina HiSeq 4000 platform (BGI, China) using 150 bp paired-end reads.

The quality control of the raw sequences was initially evaluated with the program FastQC (Babraham Bioinformatics) [30]. Reads were controlled with Cutadapt to eliminate potential ligated materials of Illumina adapters, permitting a 10% mismatch rate in the adapter sequence [31]. During this process, the programs process_radtags module and clone_filter from the software Stacks version 2.52 (http://catchenlab.life.illinois.edu/stacks/) were used to demultiplex data and trim adapter sequences and remove over-represented sequences, respectively [32]. To avoid low-quality reads with artificial bias, the following thresholds were set [33]: reads with adapter contamination were removed; reads with ≥10% unidentified nucleotides were removed; reads with >50% bases having Phred quality <20 were removed; putative duplication reads generated by PCR amplification in library construction were discarded; reads were checked for
presence of the partial 5 bp EcoRI motif (AATTC). De novo assembly was performed using software SPAdes [34]. The contigs shorter than 100 bp were removed from the assembly. To evaluate the quality of assembly, the reads involved in alignment were re-mapped onto the assembled genome using BWA program (version 0.7.17) [35].

All clean sequences were aligned to the reference assembly using BWA program. Microsatellite motif repeats of mono to hexa-nucleotide microsatellites were identified from the assembled genome using MicroSatellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). The following criteria of motifs were required to be met: at least 10 repetitive motifs for mononucleotide, 6 repetitive motifs for di-nucleotide, and 5 repetitive motifs for tri- to hexa-nucleotide. The number of microsatellite loci, repeat motifs, number of repeats, motifs sequence length, repeat type, start and end location of the repeat sequence, and microsatellite sequence, were analyzed using MISA. The software TFPGA ver. 1.3 [36] based on Nei’s [37] was used to construct Unweighted Pair Group Method with Arithmetic mean (UPGMA) dendrogram based on Nei’s [37] unbiased distance.

A total of 36 polymorphic loci were selected for validation by PCR. Polymerase chain reaction (PCR) reactions were performed in a final volume of 25 μL consisting of 1.2 μL of template DNA (100 ng), 2.5 μL of 10 × PCR buffer, 2.5 μL of Mg2+ (20 mM), 0.85 μL of dNTP mixture (2.5 mM each), 0.30 μL of each primer (10 pmol/μL), 17.35 μL of ddH2O, and 0.3 μL of 1 U/μL Taq DNA polymerase (Qiagen™, Germany). Amplified products were checked on 1% agarose gel electrophoresis and sequenced using ABI Prism 3730 automated DNA analyzer (Applied Biosystems).
The PCR amplification procedure was performed in a thermal cycler (Bio-Rad, USA) using following program: first denaturation 94 °C, 30 s, one cycle; denaturation 94 °C, 30 s, 12 cycles; a specified annealing temperature for each primer pair for 30 s, 60 s, 12 cycles; then the products were extended at 72 °C for 7 min (for annealing temperatures of each primer pair, see Table 2). The PCR products were electrophoresed on 1.2% agarose gels, stained with ethidium bromide followed by visualization under UV light. SSR loci with single bands were chosen as candidates for additional validation.

Validation of microsatellite polymorphism

For validated SSR loci, each of the selected primer sequences was validated with 120 single individual of C. aculeata from three sampling locations in Iran. From fragment sizing and reporting, GeneMapper (v4.1) was used to analyze genotype data. The negative controls with nuclease-free sterile water were included in each PCR experiment. We randomly selected 80 microsatellite loci to be amplified in nine from three different populations (three individual per population) of Capoeta aculeate for preliminary testing of all the primers. To investigate polymorphism of the loci, non-denaturing 12% polyacrylamide gel electrophoresis (PAGE) was used to screen for size variants. After the initial tests, 36 primer pairs were observed to be potentially useful and further used to test polymorphism of the loci. We carried out multiplex PCR according to the manufacturer’s instructions. We used the programs MICRO-CHECKER v.2.2.3 [38] to screen for scoring errors and allelic dropout and FREENA to estimate the frequency of null alleles in all loci [39].

For validated loci, statistics including the number of alleles (N_a), the effective number of alleles (Ne), the observed heterozygosity (H_o), expected heterozygosity (H_e) indices and inbreeding coefficient (F_Is) were calculated in GenAlex v6.5 [40]. The Hardy–Weinberg distribution (HWE) with Bonferroni correction was estimated at the population level using Arlequin v3.0 [41]. The inbreeding coefficient index (F_IS) and the polymorphism information content (PIC) were estimated using Genepop V4.7 [42].

Results

In the present study, the total GBS raw reads varied from 11.4 to 22.38 million, with total raw nucleotides ranging from 0.93 Gb to 2.96 Gb. GBS sequence analysis and microsatellite mining were performed using a modified GBS analysis workflow without using a reference genome (Fig. 2). After filtering low-quality reads and reads that lacked enzyme cutting sites by the process_radtags component of Stacks v1.39 software, the total number of clean reads kept in the libraries varied from 10.81 to 21.74 million, with an average of 15.72 million. The summary of sequencing reads generated for 30 samples of C. aculeata were listed in Table 3.

Of the 15,476 sequences, 16,706 microsatellite loci were found by MISA in C. aculeata genome sequences. The number of 926 sequences contains more than one microsatellite. Analysis of the nucleotide sequences of the 36 SSR loci revealed that dinucleotide repeats were the most abundant (6843, 51.91%), followed by trinucleotide repeats (2991, 22.69%), mononucleotide repeats (2756, 20.91%) tetrabulucleotide repeats (478, 3.62%), pentanucleotide repeats (94, 0.71%), and hexanucleotide repeats (18, 0.13%) (Fig. 3).

In the present study, the average of the allele numbers inspected at each locus varied between 1.7 for locus Ca04 to 16 for locus Ca25 (Table 3). The results showed that, the polymorphism information content (PCI) of these SSR loci varied from 0.254 to 0.888. The observed heterozygosity (H_o) per locus ranged from 0.170 to 0.881, while the expected heterozygosity (H_e) per locus was from 0.213 to 0.884. The final set of all microsatellite loci were most informative (PIC > 0.50). Across all samples, 20 SSR loci showed significant deviations from the Hardy–Weinberg equilibrium after Bonferroni correction, (p < 0.05) (Table 3).

The AMOVA implied that there was significant differences (p < 0.001) among three regions. The AMOVA also partitioned of total 23.53% genetic variation among the populations and 59.83% of the total within the individuals, indicating that much of the variation was within the individuals (Table 4). Results from the AMOVA analysis were supported by the UPGMA dendrogram for overall population differences (p < 0.001). The UPGMA dendrogram was constructed based on Nei’s unbiased genetic distance among all the examined populations of C. aculeata by using the software TFPGA. The dendrogram showed the populations of C. aculeata clustered onto two major branches (Fig. 4). The upper cluster included the Khersan and Beshar River populations while the lower cluster showed Maroun population.

Discussion

For non-model organisms, microsatellite loci were generally identified by 5 anchor PCR [43], transcriptome sequence analysis [44, 45], and microsatellite library screening [46] and development of microsatellite loci through cross-species amplification of closely related species [47]. In the present study, RAD-seq method was used for isolating microsatellite loci in C. aculeata. In contrast with conventional approaches, RAD-seq approach is considered to be faster.
Table 2 Description of 36 microsatellite markers in *Capoeta aculeata* (Valenciennes, 1844) (F = forward; R = reverse)

| Name  | Primer sequence (F and R, 5′–3′) | Allele size (bp) | Repeat motif | NA  | TM (°C) | HO  | HE  | FIS | PIC | PHWE |
|-------|----------------------------------|-----------------|-------------|-----|---------|-----|-----|-----|-----|------|
| Ca01  | F: GGC GAAC TGT GCT GAG TAAC R: GCAT GCAG ATCA ATT CAA | 202–206 | (ATT)8 | 3.8 | 54 | 0.318 | 0.432 | 0.275 | 0.635 | 0.295 |
| Ca02  | F: CTC ATCT CTT GGG ACT TGA R: CGTG CAT AT CC CAT GC | 228–238 | (GAT)9 | 4.3 | 56 | 0.854 | 0.879 | −0.028 | 0.672 | 0.000* |
| Ca03  | F: AGCT GCCT TGG GAA AGCTG R: AGCG TAG TAT GTG A GTACT | 142–164 | (CTT)4 | 6.8 | 56 | 0.722 | 0.745 | −0.048 | 0.254 | 0.926 |
| Ca04  | F: TCT GCT TAC AGC AC AT CAG A R: GTCT GTC GTCT CAG TCT | 98–106 | (CGCT)4 | 1.7 | 55 | 0.212 | 0.268 | −0.214 | 0.701 | 0.000* |
| Ca05  | F: GT TAC AGTA TG TGG CTG TCT CA R: AG GT TAC GGG TTT GAA G | 274–298 | (GAT)7 | 3.4 | 56 | 0.852 | 0.867 | 0.086 | 0.504 | 0.000* |
| Ca06  | F: GG CTAG TAAC GC GT G TGT CTA R: TG TA CAA AGC TAA CGT G | 94–110 | (AT)10 | 5 | 58 | 0.766 | 0.784 | −0.116 | 0.869 | 0.122 |
| Ca07  | F: TT GCT GAG CT CTG TAC TGA R: CGA AT AC A CGA GAG CACT A G | 100–116 | (CA)7 | 6.6 | 58 | 0.484 | 0.642 | −0.127 | 0.397 | 0.000* |
| Ca08  | F: ACC GCG CTG ATT CT GAT CG | 192–206 | (CT)8 | 2.4 | 54 | 0.214 | 0.261 | 0.115 | 0.588 | 0.048 |
| Ca09  | F: TAG GAC CAT GCT GT C TCT CTA T A R: AT CG TGA TCT CA GAC CA G | 242–274 | (TG)10 | 3.8 | 54 | 0.224 | 0.298 | 0.365 | 0.766 | 0.000* |
| Ca10  | F: TTC GA ATAG TC AT CAC AC CG | 210–242 | (AT)12 | 0.170 | 0.213 | −0.031 | 0.656 | 0.273 |
| Ca11  | F: CT TGG TCC TGG CAT TGA TG TCT C | 264–286 | (CA)8 | 3 | 52 | 0.702 | 0.746 | −0.117 | 0.828 | 0.000* |
| Ca12  | F: GTC A GTC A TCG A CTG A ACT T | 202–236 | (TG)8 | 2.0 | 58 | 0.318 | 0.389 | 0.488 | 0.771 | 0.000* |
| Ca13  | F: CAC TGG GC TGA TGG ATT AT TG A R: AT CAG GGC CAT GGG GGG | 238–276 | (AT)10 | 4.1 | 58 | 0.278 | 0.265 | 0.163 | 0.758 | 0.428 |
| Ca14  | F: ACT GC GGT GGT ATG GCT GAT CC R: TCG CC GGC ACT AT CACT AGG GA | 210–224 | (TC)7 | 4.2 | 50 | 0.811 | 0.884 | 0.436 | 0.857 | 0.000* |
| Ca15  | F: GA CT A GTC GGC GCC AGC CA GTAC R: AT CGT TAG ATG GCT ATT GC | 246–282 | (AA)6 | 6.0 | 50 | 0.633 | 0.679 | −0.130 | 0.835 | 0.000* |
| Ca16  | F: TG CT C GT GA TG GT GT GTA TAC R: GTA AG TCT CAC TTT G CT TGC | 158–184 | (ATC)6 | 8.4 | 52 | 0.593 | 0.489 | 0.015 | 0.747 | 0.178 |
| Ca17  | F: TAG C A GTG TGA T A C T G | 86–96 | (TG)6 | 2.6 | 54 | 0.571 | 0.743 | 0.228 | 0.825 | 0.000* |
| Ca18  | F: GGT TGC ATC GT A CTG A CA A AT C R: GGC TCG AT A GT GCA TAC TTA C | 254–278 | (TG)6 | 3.8 | 54 | 0.774 | 0.627 | 0.286 | 0.867 | 0.387 |
| Ca19  | F: TTC CG TAC AT C TAC AGC CG | 112–128 | (TCC)8 | 10 | 58 | 0.633 | 0.649 | 0.523 | 0.341 | 0.000* |
| Ca20  | F: GT CT GCA TAC AGC AT CCG GT C R: TCA CT G TG ACT C TCT C TCT C TCT | 95–106 | (AG)9 | 4.4 | 60 | 0.562 | 0.476 | −0.123 | 0.643 | 0.188 |
| Ca21  | F: ACT TGA GT TAC TCG A CT G G T A R: TCG CAT AG CT TCT AT GC | 180–204 | (TC)8 | 4.0 | 58 | 0.419 | 0.664 | 0.606 | 0.830 | 0.000* |
| Ca22  | F: CG A TA TGG GA AG C TG G GT TTT G T T T T | 227–263 | (ATAC)6 | 3.1 | 50 | 0.318 | 0.390 | −0.044 | 0.682 | 0.028 |
| Ca23  | F: ACT CG C TA A T C A C G A T C G AT C R: TGT CAT G AC T C T A G T | 158–194 | (TC)12 | 5.4 | 52 | 0.579 | 0.805 | −0.171 | 0.797 | 0.000* |
| Ca24  | F: GC A T G A T C T A C T GC GG A G TA R: GT G A C AT G C A G C T | 102–136 | (TATC)14 | 4.0 | 52 | 0.290 | 0.478 | 0.345 | 0.692 | 0.000* |
| Ca25  | F: AG T A G C A AG G AT T C AT G T A R: TCG GTG ACA T G C A G | 312–374 | (TA)5 | 16 | 53 | 0.370 | 0.584 | 0.244 | 0.568 | 0.000* |
Table 2 (continued)

| Name  | Primer sequence (F and R, 5′–3′) | Allele size (bp) | Repeat motif | N_A | TM (°C) | H_O | H_E | F_IS | PIC | PHWE |
|-------|----------------------------------|------------------|--------------|-----|---------|-----|-----|------|-----|------|
| Ca26  | F: TTGCCATTAATCGTAGCGACA R: GCTACTAAACCGCTATCGTG | 224–238 | (AT)6 | 2.6 | 57 | 0.675 | 0.785 | 0.326 | 0.731 | 0.005 |
| Ca27  | F: CGTCCCGCGCTTTACATCTTGGA R: GCTCGTGCTCCTCGCCATGTCG | 124–154 | (AT)6 | 4.1 | 51 | 0.325 | 0.843 | 0.494 | 0.745 | 0.000* |
| Ca28  | F: GATTCTCTAAAGTCACCTTTGTGT R: TAGACACTTGACATTCTATCG | 114–140 | (GT)5 | 14.0 | 57 | 0.525 | 0.713 | 0.283 | 0.634 | 0.000* |
| Ca29  | F: GGTGTGCAATCTTCATACTAGC R: CTTGTGTCGACATTGCTAGGGC | 192–210 | (AT)5 | 4.2 | 58 | 0.472 | 0.624 | −0.283 | 0.888 | 0.000* |
| Ca30  | F: TGGTCGGAGTTAAGCTAGGC R: GAGTACGTCGTTGGCTACTCAA | 134–174 | (TA)5 | 2.0 | 54 | 0.551 | 0.643 | −0.215 | 0.582 | 0.094 |
| Ca31  | F: GGCTAATACCTAGAGTTATG R: GTCGGGATATGCATGTCGTA | 142–148 | (CT)8 | 2.2 | 53 | 0.188 | 0.680 | −0.549 | 0.265 | 0.343 |
| Ca32  | F: AGGTGATGCACAAATCGGGTCGC R: TAGCTCAGCTCACCCTCTAGAG | 162–168 | (TA)14 | 4.4 | 55 | 0.207 | 0.642 | −0.074 | 0.468 | 0.288 |
| Ca33  | F: TCAGAGATGCTCGGATACGTCA R: ACGCTCTGTAGGAAGCAGG | 110–131 | (TG)6 | 5.8 | 56 | 0.625 | 0.685 | −0.495 | 0.726 | 0.009 |
| Ca34  | F: GCATACGCAGACTGTATTCGTCG R: AGCTAGCTGATGCTATCAGTC | 121–154 | (AATC)4 | 9.2 | 55 | 0.567 | 0.802 | 0.368 | 0.387 | 0.000* |
| Ca35  | F: GTCTCACTCATCCATGGTGTCG R: GAACGGATACACTCTTCTGTC | 212–224 | (AG)9 | 4.2 | 56 | 0.438 | 0.442 | 0.166 | 0.368 | 0.098 |
| Ca36  | F: AGTCTCTAGCTCATCCTCAGCTAG R: CCGAGTTAAATGAGACGATAT | 264–268 | (TG)5 | 4.4 | 57 | 0.231 | 0.587 | 0.216 | 0.467 | 0.000* |

H_O observed heterozygosity, H_E expected heterozygosity, N_A number of allele, F-IS inbreeding coefficient, PIC polymorphism information content, PHWE p-values for Hardy–Weinberg equilibrium

*Indicates significant deviation from Hardy–Weinberg Equilibrium after Bonferroni correction

Fig. 2 Bioinformatic workflow for GBS-based SSR sequence identification

Genotyping by sequence
471752360 clean reads

Unique sequence
17,260,075

Sequence catalogs
2758956

Sequence containing SSRs region
198056

All sample read alignment with sequence containing SSRs region

Allel frequency table
1956 polymorphic SSRs
Table 3 Summary of sequencing reads generated for each samples of *Capoeta aculeata*

| Name    | Raw reads (million) | Raw nucleotides (Gb) | Clean reads (million) | Genome coverage | Unique reads containing SSRs |
|---------|---------------------|----------------------|-----------------------|----------------|-----------------------------|
| Ca1-kh  | 11.42               | 1.25                 | 10.81                 | 0.62           | 641,025                     |
| Ca2-kh  | 14.32               | 1.58                 | 13.35                 | 0.68           | 578,246                     |
| Ca3-kh  | 15.65               | 1.45                 | 14.56                 | 0.64           | 625,268                     |
| Ca4-kh  | 13.44               | 1.42                 | 11.30                 | 0.59           | 396,795                     |
| Ca5-kh  | 14.65               | 1.74                 | 12.48                 | 0.74           | 605,436                     |
| Ca6-kh  | 17.21               | 1.69                 | 16.19                 | 0.69           | 508,294                     |
| Ca7-kh  | 14.24               | 1.42                 | 13.28                 | 0.57           | 512,346                     |
| Ca8-kh  | 11.86               | 1.24                 | 10.67                 | 0.52           | 482,247                     |
| Ca9-kh  | 13.12               | 1.43                 | 12.49                 | 0.54           | 514,345                     |
| Ca10-kh | 15.49               | 1.51                 | 14.44                 | 0.63           | 758,478                     |
| Ca11-Be | 11.4                | 1.28                 | 11.21                 | 0.56           | 524,635                     |
| Ca12-Be | 16.52               | 2.35                 | 16.22                 | 0.85           | 598,075                     |
| Ca13-Be | 18.36               | 2.35                 | 18.12                 | 0.81           | 605,462                     |
| Ca14-Be | 16.45               | 2.96                 | 15.96                 | 0.74           | 635,153                     |
| Ca15-Be | 18.24               | 2.12                 | 17.47                 | 0.69           | 563,508                     |
| Ca16-Be | 14.65               | 2.1                  | 13.35                 | 0.54           | 532,864                     |
| Ca17-Be | 20.35               | 2.3                  | 20.01                 | 0.92           | 478,652                     |
| Ca18-Be | 16.25               | 1.6                  | 15.87                 | 0.68           | 515,728                     |
| Ca19-Be | 19.34               | 2.25                 | 18.65                 | 1.14           | 496,254                     |
| Ca20-Be | 21.34               | 3.5                  | 20.04                 | 0.74           | 630,542                     |
| Ca21-Ma | 17.98               | 1.82                 | 17.13                 | 0.86           | 563,435                     |
| Ca22-Ma | 18.27               | 2.47                 | 17.62                 | 0.54           | 635,028                     |
| Ca23-Ma | 19.12               | 2.35                 | 18.97                 | 0.98           | 543,759                     |
| Ca24-Ma | 14.22               | 1.85                 | 13.35                 | 0.87           | 431,078                     |
| Ca25-Ma | 16.35               | 1.5                  | 15.25                 | 1.02           | 486,381                     |
| Ca26-Ma | 19.28               | 1.25                 | 18.26                 | 0.69           | 658,582                     |
| Ca27-Ma | 18.47               | 0.93                 | 17.98                 | 0.94           | 546,865                     |
| Ca28-Ma | 19.25               | 2.84                 | 19.1                  | 0.75           | 865,146                     |
| Ca29-Ma | 16.1                | 2.37                 | 15.88                 | 0.64           | 561,353                     |
| Ca30-Ma | 22.38               | 1.81                 | 21.74                 | 0.99           | 765,095                     |
| Avg     | 16.52               | 1.85                 | 15.72                 | 0.73           | 575,336                     |
| Max     | 22.38               | 2.96                 | 21.74                 | 1.14           | 865,146                     |
| Min     | 11.4                | 0.93                 | 10.81                 | 0.52           | 431,078                     |
| Total   | 495.72              | 55.61                | 471.75                | 22.17          | 17,260,075                  |

Fig. 3 Frequency of SSR loci identified in *Capoeta aculeata*
and profitable. Furthermore, this method can produce a large number of microsatellite markers for one time [17, 25]. Population structure genetic is very important for the sustainability of many species. For example for Labeo bata [48], blunt snout bream Megalobrama amblycephala [49], sea cucumber Holothuria scabra [50], red snapper Lutjanus campechanus [51], pearl oyster Pinctada fucata [52], catfish, Rhamdia sp. [53], yellow catfish Pseudobagrus fulvidraco [54], Persian sturgeon (Acipenser persicus) [55], narrow-clawed crayfish Pontastacus leptodactylus [27] and fan mussel Pinna nobilis [56]. Conservation management plans with no prior information of the genetic background could result in disturbance to the genetic structure with adverse effects on the gene pools of wild populations [18]. Until now, the microsatellite loci have not been carefully developed for C. aculeata which has posed a serious obstacle to conservation and management of this species [13]. The current study detected 16,706 microsatellite markers in C. aculeata genome using MISA, which accounted for 2.12% of the total genome sequence. The relative abundance of microsatellite sequences was calculated at 1.16 loci per kb of C. aculeata genomes. In general, the frequencies of microsatellite loci are expected to decrease with increasing repeat length due to longer repeats have a higher possibility of being mutated [57], and this tendency has been identified in many organisms [58–60]. According to Nei’s unbiased genetic distance among all the investigated populations of C. aculeata, the UPGMA dendrogram clustered the populations onto two major branches indicating their genetic relatedness partially following their geographic distribution. The populations lying in the same cluster with restricted genetic distance despite larger geographic distance indicate the stock management structure of fisheries department in the province.

### Conclusion

This study is the first report of SSR loci in C. aculeata developed using RAD-seq. In the present study we observed SSR loci for Capoeta aculeata from contigs assembled from GBS data based on the Illumina HiSeq platform. From the identified contigs, 36 polymorphic SSR markers were characterized. Moreover, we also indicated that NGS data is a useful method for genotyping of this species. Therefore these SSR markers would provide an invaluable resource for population genetics and natural resource conservation in C. aculeata [57–60].

### Acknowledgements

This work was supported by the Research Fund for the department of Fisheries of the Islamic Azad University of Tehran, Iran.

### Author contributions

HG made the literature, sample collection, material preparation. SPHS has major contribution to the preliminary study design and interpretation of data. The first draft of the manuscript was written by HAA. SN designed the experiment, institutional support and supervision of the preliminary experiment and made the data analysis. MSM has significant contributions to the laboratory work and interpretation of the results. All authors read and approved the final manuscript.

### Data availability

The datasets generated are available from the corresponding author upon request.

### Declarations

#### Conflict of interest

The authors declare that they have no conflict of interest.

#### Ethical approval

The main research project researcher of this work was approved for animal ethics by Iranian Fisheries Sciences Research Institute (IFSRI) for scientific purposes development, Iran. All applicable guidelines for the care and use of animals were followed by the authors. The experiment was approved for Animal Welfare by the Institutional Animal Welfare Committee of Agricultural Research, Education and Extension Organization (AREEO), Iran.

#### Informed consent

This research does not involve humans and therefore informed consents are not applicable.

#### Consent to participate (ethics)

Informed consent was obtained from all individual participants included in the study.

| Source of variation | df | Sum of squares | Variance components | Percentage of variation |
|---------------------|----|----------------|---------------------|-------------------------|
| Among populations   | 2  | 735.495        | 2.15635             | 23.53                   |
| Among individuals within populations | 114 | 1135.274 | 1.38695 | 16.64 |
| within individuals  | 117 | 968           | 5.69352             | 59.83                   |
| Total               | 283 | 2838.769      | 9.23682             | 100.00                  |

**Fig. 4** UPGMA dendrogram based on the genetic distance in the populations of C. aculeata
References

1. Zarein H, Esmaeili HR, Freyhof J (2016) Capoeta amanansimis, a new species from the Minab and Hasan Langhi River drainages in Iran (Teleostei: Cyprinidae). Zootaxa 4083:126–142
2. Bektas Y, Turan D, Aksu I, Ciftci Y, Eroglu O, Kalayci G, Belduz AO (2017) Molecular phylogeny of the genus Capoeta (Teleostei: Cyprinidae) in Anatolia. Turkey Biochem Syst Ecol 70:80–94
3. Freyhof J, Esmaeeli HR, Sayyadzadeh G, Geiger M (2014) Review of the crested loaches of the genus Paracobitis from Iran and Iraq with the description of four new species (Teleostei: Nemacheilidae). Ichthyol Explor Freshw 25(1):11–38
4. Nazari S, Paknejad H, Jalali A, Khorshidi Z (2021) Molecular genetic diversity of the cyprinid fish Craterocephalus dimidiatus from China, Russia, South Korea and Japan as revealed by microsatellite markers. Mar Ecol 30:455–461
5. Parmaksiz A, Eksi E (2017) Genetic diversity of the cyprinid fish Capoeta trutta (Heckel, 1843) populations from Euphrates and Tigris rivers in Turkey based on mtDNA COI sequences. Indian J Fish 64(1):18–22
6. Sunnucks P (2000) Efficient genetic markers for population biology. Trends Ecol Evol 15:199–203
7. Zane L, Bargelloni L, Patarello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11:1–16
8. Chang Y, Feng Z, Yu J, Ding J (2009) Genetic variability analysis in populations of Veneridae (Decapoda: Ostreidae) by microsatellite markers. Mar Ecol Prog Ser 388:135–148
9. Xue D, Zhang T, Liu J (2014) Microsatellite evidence for high frequency of multiple paternity in the marine gastropod Rapana venosa. PLoS ONE. https://doi.org/10.1371/journal.pone.0086508
10. Montanari S, Perchepied L, Renault D, Frijters L, Velasco R, Horner M, Gardiner SE, Chagné D, Bus VGM, Durel CE, Malnoy M (2016) A QTL detected in an inter specific pear population confers stable fire blight resistance across different environments and genetic backgrounds. Mol Breeding 36:1–16
11. Jia SW, Liu P, Li J, Li JT, Pan LQ (2013) Isolation and characterization of polymorphic microsatellite loci in the ridgeback white prawn Exopalaemon carinicauda. Genet Mol Res 12:2816–2820
12. Perina A, Gonzalez-Tizn AM, Vizzano A, Gonzalez-Ortegon E, Martinez-Lage A (2016) Isolation and characterization of 20 polymorphic microsatellite loci in Palaeonidae species by 454 pyrosequencing. Conser Genet Resour 8:169–196
13. Gandomkar H, Shekarabi SPH, Abdolhay HA, Nazari S, Mehrjan MS (2020) Genetic structure of the Capoeta aculeata populations inferred from microsatellite DNA loci. Biodiversitas 21:4565–4570. https://doi.org/10.3057/biodiv/d211014
14. Dutta N, Singh RK, Mohindra V, Pathak A, Kumar R, Sah P, Mandal S, Kaur G, Lal KK (2019) Microsatellite marker set for genetic diversity assessment of primitive Chitala chitala (Hamilton, 1822) derived through SMRT sequencing technology. Mol Biol Rep 46(1):41–49
15. Qiu B, Fang S, Ikhwanuddin M, Wong L, Ma H (2020) Genome survey and development of polymorphic microsatellite loci for Sillago sihama based on Illumina sequencing technology. Mol Biol Rep 47(4):3011–3017
16. Feng N, Ma H, Ma C et al (2014) Characterization of 40 single nucleotide polymorphism (SNP) via T m-shift assay in the mud crab (Scylla paramamosain). Mol Biol Rep 41:5467–5471. https://doi.org/10.1007/s11033-014-3420-2
17. Xiao M, Hu Q, Zhao Y, Bao F, Cui F, Zheng R (2018) Development of 36 SNP markers in Ophiocolasus argus Cantor base on high throughput sequencing. Conserv Genet Resour 10(1):35–38. https://doi.org/10.1007/s12686-017-0757-6
18. Khoshkholgh M, Nazari S (2020) Characterization of single nucleotide polymorphism markers for the narrow-clawed crayfish Pontastius leptodactylus (Escholtz, 1823) based on RAD sequencing. Conserv Genet Resour. https://doi.org/10.1007/s12686-020-01154-8
19. Ewers-Saucedo C, Zardus JD, Wares JP (2016) Microsatellite loci discovery from next-generation sequencing data and loci characterization in the epizooic barnacle Chelonibia testudinaria (Linnaeus, 1758). PeerJ. https://doi.org/10.7717/peerj.2019
20. Wei N, Bemmelns JB, Dick CW (2014) The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. Mol Ecol Resour 14:953–965
21. Davey J, Hohenlohe PA, Etter PD, Boone JQ, Catches JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510
22. Zhou J, Zhou B, Li Q et al (2017) Isolation and characterization of 33 EST-SNP markers in Schizothorax pretani. Conserv Genet Resour 10:205–207. https://doi.org/10.1007/s12686-017-0799-9
23. Yu L, Bai J, Cao T et al (2014) Genetic variability and relationships among six grass carp Ctenopharyngodon idella populations in China estimated using EST-SNP markers. Fish Sci 80:475–481. https://doi.org/10.1007/s12686-014-0709-y
24. Luo H, Ye H, Xiao S-J et al (2016) Development of SNP markers associated with immune-related genes of Schizothorax pretani. Conserv Genet Resour 8:223–226. https://doi.org/10.1007/s12686-016-0539-6
25. Pang X, Zhao L, Liu J, Guo X (2018) Development of SNP markers for Xenopsris argentea based on transcriptomics. Conserv Genet Resour 10(4):679–684. https://doi.org/10.1007/s12686-017-0900-4
26. Nazari S, Jafari V, Pourkazemi M, Kolangi Miandare H, Abdolah H (2016) Association between myostatin gene (MSTN-1) polymorphism and growth traits in domesticated rainbow trout (Oncorhynchus mykiss). Agri Gene 1:109–115
27. Khoshkholgh M, Nazari S (2019) The genetic diversity and differentiation of narrow-clawed crayfish Pontastius leptodactylus (Escholtz, 1823) (Decapoda: Astacidae: Astacidae) in the Caspian Sea Basin, Iran as determined with mitochondrial and microsatellite DNA markers. J Crust Biol 39(2):112–120
28. Wang S, Liu P, Lv J, Li Y, Cheng T, Zhang L (2016) Serial sequencing of iso-length pad tags for cost-efficient genome-wide profiling of genetic and epigenetic variations. Nat Protoc 11(11):2189–2200. https://doi.org/10.1038/nprot.2016.133
29. Blian L, Liu C, Liu G, Chen S, Zhang L, Ge J, Li F, Tan J (2018) SNP discovery in spotted halibut (Verasper variegatus) using restriction site-associated DNA sequencing (RAD-seq). Conserv Genet Resour 10(5):409–413
30. Andrews S (2016) FastQC: a quality control tool for high throughput sequence data. Version 0.11.5. Babraham Institute, Cambridge, UK. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
31. Martin M (2011) Sequencing reads. EMBNet J 17:10–12
32. Catches JP, Hohenlohe SB, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22(11):3124–3140
33. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. J Bioinf Comput Biol 22(13):1658–1659
