Mol Genet Genomics (2017) 292:1165–1174
DOI 10.1007/s00438-017-1337-x

METHODS PAPER

A rapid and cost-effective approach for the development of polymorphic microsatellites in non-model species using paired-end RAD sequencing

Dong-Xiu Xue[1,2] · **Yu-Long Li**[1,2] · **Jin-Xian Liu**[1,2]

Received: 24 March 2017 / Accepted: 14 June 2017 / Published online: 20 June 2017

Abstract As one of the most informative and versatile DNA-based markers, microsatellites have been widely used in population and conservation genetic studies. However, the development of microsatellites has traditionally been laborious, time-consuming, and expensive. In the present study, a rapid and cost-effective "RAD-seq-Assembly-Microsatellite" approach was developed to identify abundant microsatellite markers in non-model species using the roughskin sculpin *Trachidermus fasciatus* as a representative. Overlapping paired-end Illumina reads generated by restriction-site-associated DNA sequencing (RAD-seq) were clustered based on the similarity of reads containing the restriction enzyme recognition site and then assembled into contigs, which were used for microsatellite discovery and primer design. A total of 121,750 RAD contigs were generated with a mean length of 522 bp, and 19,782 contigs contained microsatellite motifs. A total of 156,150 primer pairs were successfully designed based on 16,497 contigs containing priming sites. Experimental validation of 52 randomly selected microsatellite loci demonstrated that 45 (86.54%) loci were successfully amplified and polymorphic in two geographically isolated populations of *T. fasciatus*. Compared with traditional approaches based on DNA cloning and other approaches based on next-generation sequencing, our newly developed approach could yield thousands of microsatellite loci with much higher successful amplification rate and lower costs, especially for non-model species with shallow background of genomic information. The "RAD-seq-Assembly-Microsatellite" approach holds great promise for microsatellite development in future ecological and evolutionary studies of non-model species.

Keywords Microsatellite · Next generation RAD sequencing · Population genetics · Non-model species · *Trachidermus fasciatus*

Communicated by S. Hohmann.

Dong-Xiu Xue and Yu-Long Li contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1337-x) contains supplementary material, which is available to authorized users.

✉ Jin-Xian Liu
jinxianliu@gmail.com

[1] CAS Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, 7 Nanhai Road, Qingdao 266071, Shandong, China

[2] Laboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266071, Shandong, China

Introduction

Microsatellites, also known as simple sequence repeats (SSRs), are tandem repeats of one to six nucleotides in DNA sequences (Oliveira et al. 2006). Given their extensive distribution in genome, high polymorphism, codominant inheritance and high success amplification rate, microsatellites have been one of the most powerful and valuable molecular tools in many research areas, such as population genetics, conservation genetics, genome mapping, parentage analysis, and quantitative trait loci identification (Chang et al. 2009; Montanari et al. 2016; Xue et al. 2014). Despite advances in the achievement of single nucleotide polymorphism (SNP) data, microsatellites are still useful and more easily accessible for many studies such as those involving genetic diversity monitoring for a long period of stock management, and breeding or pedigree estimation (Hodel et al. 2016; Minegishi et al. 2015; Stabile et al. 2016; Weinman et al. 2015; Zalapa et al. 2012). A major limitation to the usage of microsatellites is that traditional methods for microsatellites development, such as an enriched library followed by cloning and Sanger sequencing, were labor-intensive, time-consuming, and expensive (Glenn and Schable 2005; Zane et al. 2002). Furthermore, microsatellites had to be developed de novo for every species under study, as cross-amplification from congeneric species is not generally feasible (Schoebel et al. 2013). Therefore, rapid and cost-effective methods for microsatellite development are urgently needed for population management and conservation of non-model species.

Next generation sequencing (NGS) can produce large amount of sequences, from which numerous genome-wide and gene-based microsatellite loci could be isolated and developed (Zalapa et al. 2012). So far, studies using NGS to develop microsatellite loci have largely rely on the Roche 454 and Illumina sequencing platforms (Hodel et al. 2016; Minegishi et al. 2015; Schoebel et al. 2013; Zalapa et al. 2012). Since read length is an important factor that affects the possibility to discover microsatellites and design primers, the 454 sequencing platform was used extensively for microsatellite development (Hodel et al. 2016; Mastretta-Yanes et al. 2015). However, 454 is less cost-effective than Illumina on a per-megabase basis (Glenn 2011; Zalapa et al. 2012). Moreover, Roche have discontinued the use of the 454 instrument since 2016. Currently, projects of microsatellite discovery largely focused their efforts on Illumina platforms. However, the short read lengths obtained with Illumina platform limited its utility for microsatellite development, because most reads did not have enough flanking sequences for primer design. To improve the efficiency of microsatellite development using Illumina reads, sequence assembly should be performed to create longer DNA sequences or contigs before microsatellite discovery.

One cost-efficient and practical strategy to develop microsatellite markers using NGS technologies is the sequencing of a reduced representation genomic library (Bonatelli et al. 2015). Restriction site-associated DNA sequencing (RAD-seq) is a useful approach to create reduced representation genomic libraries and provide sequence data adjacent to restriction enzyme recognition sites (Davey et al. 2011; Hohenlohe et al. 2013). RAD-seq incorporates a random shearing step in library preparation, which can be modified to generate overlapping paired reads. The reads of a single RAD locus generated by traditional RAD-seq technology could be firstly clustered using the similarity of the first reads with the restriction enzyme recognition site, and then the overlapping paired-end reads allowed local assembly of contigs containing both the forward and reverse reads of each pair (Hohenlohe et al. 2013), which could improve accuracy and quality of the assembled contigs, and therefore improve the success rate of microsatellite development. The roughskin sculpin *Trachidermus fasciatus* Heckel (Scorpaeniformes: Cottidae), is a small, benthic, carnivorous, and catadromous fish species with a native distribution in Northwestern Pacific distribution (Onikura et al. 2002; Wang 1999). In the past decades, it has experienced severe population declines in China, probably due to degradation of habitats, water pollution and dam construction (Wang and Cheng 2010). However, only a few molecular genetic resources are publicly available for roughskin sculpin (Xu et al. 2008; Zeng et al. 2012), and the use of microsatellite markers in conservation genetic studies and maker-assisted selection was limited (Li et al. 2016b).

In the present study, a "RAD-seq-Assembly-Microsatellite" approach was developed and applied in the roughskin sculpin as a representative of non-model species, for which limited genetic data were available. To improve the success rate of microsatellite development in a simple, fast, and economic way, the advantages offered by the traditional RAD-seq technology coupled with fast and efficient bioinformatic tools for reads assembly, microsatellite isolation and primer design were explored in this approach. The essence of the approach is to generate enough long contiguous sequences of high quality to overcome technical limitations introduced by short read lengths and to isolate abundant microsatellite loci. Briefly, genomic DNA of a roughskin sculpin individual was sequenced using the overlapping paired-end RAD-seq protocol, and the generated reads were sorted according to RAD loci and locally assembled to achieve longer contiguous sequences. Then microsatellite sequences in the assembled contigs were searched and primer pairs were designed. Finally, 52 microsatellite loci were randomly selected and validated in two roughskin sculpin populations based on PCR amplification and genotyping. The newly developed rapid and cost-effective approach would be of particular advantage for the isolation and characterization of sufficient microsatellite loci for ecological and evolutionary studies of non-model species.

Materials and methods

Sampling and genomic DNA extraction

A total of 48 individuals were collected from two geographic sites in China: 24 individuals from Dandong, Liaoning Province (39°46′N, 124°20′E) in May 2014, and 24 individuals from Fuyang, Zhejiang Province (30°03′N, 119°58′E) in January 2014. Muscle tissue were preserved in 95% ethanol. Genomic DNA was extracted using the standard phenol–chloroform extraction protocol, and checked using 1% agarose electrophoresis and Nanodrop 2000c spectrophotometer.

Library preparation and RAD tag sequencing

Approximately 1 µg of genomic DNA extracted from a single individual of Fuyang was digested with restriction enzyme *Eco*RI. The digested products were ligated to a modified Illumina P1 adapter containing individual-specific index sequences of 6 bp for sample tracking. The total genomic DNA samples were then randomly sheared to an average size of 500 bp, and fragments with insert size spanning 200–600 bp were isolated using a MinElute Gel Extraction kit (Qiagen). An "A" base overhangs were added to the 3′ ends of the blunt DNA fragments, and then a modified P2 adapter containing a 3′ dT overhang was ligated onto the ends of DNA fragments with 3′ dA overhangs. Finally, the library was enriched by high-fidelity PCR amplification, preparing RAD tags that contain both adaptors for paired-end (2 × 125 bp) sequencing on an Illumina Hi-Seq 2500 platform at Novogene in Tianjin.

RAD data assembly and assessment

Illumina raw reads were quality-filtered, and PCR duplicates were removed by "clone_filter" in STACKS (version 1.32) (Catchen et al. 2013). The first reads with restriction enzyme recognition sites were sent to STACKS to identify RAD loci. The minimum depth of stacks was set to 10, and the number of mismatches allowed between stacks was set to 3 to maintain the true alleles from paralogues. Deleveraging and removal algorithms were turned on to filter out highly repetitive loci. Finally, the second reads corresponding to each RAD locus were collected into separate files using a modified version of "sort_read_pairs.pl" in STACKS. The reads for each locus were locally assembled by CAP3, which is a DNA sequence assembly program based on overlap-layout-consensus methods (Huang and Madan 1999). The assembly was performed by a custom developed multi-threading Perl scripts CP3_Opti.pl (available at https://github.com/lyl8086/RAD_SSR) according to an optimized assembly approach. Firstly, the second reads for each RAD loci identified by the first reads were locally assembled into contigs. Secondly, the assembled contigs of the second reads were merged with the corresponding consensus sequences of the first reads for each RAD locus. Thirdly, a final assembly was performed on each RAD loci to generate the final assembled RAD reference. In general, the overlapping paired reads generated by RAD-seq are staggered over a local genome location, these reads can be locally assembled into high-quality contigs, which are up to 1 kb depending on the strategy of size selection in the library preparation. The longer contigs thus provided sufficient sequences for the downstream microsatellites discovery and primer design.

In order to check the quality of the assembled RAD reference, the paired reads used for assembling were mapped back to the reference by BWA 0.7.12 (Li and Durbin 2009). BWA "mem" (Li 2013) was used to generate SAM file, the parameters were set to default except for the minimum seed length of 32. The SAM file was processed by SAMTOOLS 1.3.1 (Li et al. 2009) to check the overall coverage, the number of mapped reads and the depth. To further improve the quality of the assembled RAD reference, only contigs with properly mapped read pairs (paired reads mapped in right direction with proper insert size given by the aligner) and a minimum mapping quality of 20 were retained. Soft or hard clipped reads, secondly aligned reads, and reads with the SAM tags of "XA" or "SA" were also removed. The generated high-quality contigs were then used for downstream microsatellites discovery and primer design.

Microsatellites searching and primer design

QDD 3.1.2 (Meglécz et al. 2010) was chosen for microsatellite discovery and primer design. The program was run in a local Galaxy (Afgan et al. 2016) platform with default parameters. Microsatellite was defined as pure or compound tandem repeats of di- to hexa-nucleotide motif with at least five uninterrupted repeats. To improve the success rate, primers were selected based on the following five criteria: (1) select one primer for each locus; (2) select pure microsatellites with repeat number great than 5; (3) select primers that were only in design category A; (4) remove primers with alignment score greater than 10; and (5) select primers that were away from the target microsatellite (>10 bp).

Microsatellite genotyping and polymorphism survey

A total of 52 primer pairs were randomly selected for laboratory verification. Initial testing for PCR amplification used two individuals from Fuyang. A M13-tail (5′-GGAAACAGCTATGACCATG-3′) was added on the 5′ end of each forward primer. PCR amplification were performed in a total volume of 10 µL containing 10 ng genomic DNA, 1× PCRmix (Dongsheng Biotech Co., China) and 0.2 µM each primer, using the following cycling conditions: (1) initial activation step for 5 min at 95 °C; (2) 35 cycles of denaturation at 95 °C for 20 s, annealing at 52 °C for 30 s and extension at 72 °C for 30 s; and (3) a final extension of 5 min at 72 °C. The PCR products were electrophoresed on a 1.5% agarose gel and only primers that produced specific products were further evaluated using an initial set of eight individuals. PCR amplification were carried out in a final volume of 10 μL containing 10 ng genomic DNA, 1× PCRmix (Dongsheng Biotech Co., China), 0.02 μM forward primer, 0.2 μM reverse primer, and 0.2 μM of M13-tail primer that was fluorescently labeled with FAM, HEX or TAMRA. The PCR amplification program was the same as mentioned above. Fluorescently labeled PCR fragments were electrophoresed on an ABI 3730xl automated sequencer (Applied Biosystems) with the GS-500 size standard. Allele calling was performed using GeneMarker (SoftGenetics, State College, USA). The final scoring was manually checked to minimize genotyping errors. Finally, polymorphism of microsatellite loci screened out by the above two steps were checked in 48 individuals from Fuyang and Dandong.

Genetic diversity indices for each loci and population including observed heterozygosity (H_O), expected heterozygosity (H_E) and polymorphism information content (PIC) were calculated using the Excel Microsatellite Toolkit (Park 2001). The number of alleles (*N*a), allelic richness (A_R) and inbreeding coefficient (F_{IS}) was calculated using FSTAT 2.9.3 (Goudet 2001). Deviations from Hardy–Weinberg equilibrium and genotypic linkage equilibrium were tested with Genepop 4.5.1 (Rousset 2008). The significance tests were estimated by the Markov chain Monte Carlo (MCMC) method (10,000 dememorization steps, 1000 batches of 10,000 iterations). Micro-checker 2.2.3 (van Oosterhout et al. 2004) was used to test for the presence of null alleles. A standard Bonferroni correction was used for all above significance levels of tests.

Results

RAD sequencing, filtering and assembly

A total of 33.5 million raw paired reads were obtained, and 25.1 million clean paired reads were retained after quality filtering and removing PCR duplications. A total of 137,409 loci identified by STACKS were exported into separate fasta files for local assembly. CAP3 assembled a total of 127,864 contigs with a mean length of 517 bp and N50 of 543 bp. About 20.8 million reads could be mapped back to the assembled contigs, and 94% of these were properly paired. After retaining contigs with properly paired reads and a minimum mapping quality of 20, and removing clipping and other possible spurious reads, a total of 121,750 contigs were retained as the final assembled RAD reference for microsatellites discovery (Online Resource 1). The final assembled RAD reference had a mean length of 522 bp and GC content of 41.59%.

Microsatellite isolation and characterization

A total of 19,782 contigs possessing microsatellite motifs were identified. For 16,497 contigs that contained priming sites for microsatellite loci, the types of the microsatellites in the target region were variable. The number of the pure microsatellites was 12,127, while the number of the compound microsatellites was 3242. Finally, a total of 156,150 primer pairs sets were successfully designed (Online Resource 2). Using one primer pair for each locus, pure microsatellites, design category A, PCR primer align score ≤10, and minimum primer target distance >10 bp, a total of 1854 primer pairs were retained (Table 1). These 1854 microsatellite motifs included 1536 di- (82.85%), 262 tri- (14.13%), 49 tetra- (2.64%), 5 penta- (0.27%) and 2 hexa- (0.11%) nucleotide repeats, of which the repeats number ranged from 5 to 41.

Of the 52 primer pairs randomly selected, 48 primer pairs produced clear and specific amplification products of the expected size by being screened in 1.5% agarose electrophoresis in two individuals, and were subsequently used for evaluation with capillary to test genotyping in eight individuals. Finally, a set of 45 microsatellite loci were used to evaluate polymorphism in 48 individuals from the two populations, and a total of 618 alleles were detected (Table 2). The number of alleles per locus ranged from 3 to 29, and the expected (H_E) and observed (H_O) heterozygosity ranged from 0.3510 to 0.9800 and from 0.2080 to 1.0000, respectively. The polymorphism information content (PIC) ranged from 0.3070 to 0.9590. No linkage disequilibrium was detected between microsatellite loci. Significant deviation from Hardy–Weinberg equilibrium was observed in four loci (tfa 28, tfa 32, tfa 36 and tfa 57), of which two loci (tfa 32 and tfa 57) were significant in both tested populations (Table 2). Analyses using micro-checker indicated the presence of null alleles at the same four loci.

Discussion

The present study, with the roughskin sculpin as a representative of non-model species, developed a rapid

Table 1 Summary of the number of primer pairs designed from the 16,497 assembled RAD contigs for *Trachidermus fasciatus*

Filtering steps	Number of primer pairs
Total primer pairs designed	156,150
One primer for each locus	16,497
Pure microsatellites	12,127
Design category A	2735
PCR primer align score ≤10	27,11
Min primer target distance >10 bp	1854

Table 2 Characterization of 45 microsatellite loci validated in populations of *Trachidermus fasciatus* from Fuyang and Dandong

Locus	Repeat motif	Primer sequence (5′-3′)	Size range (bp)	*Na*		H_E		H_O		*PIC*	
				Fuyang[a]	Dandong[b]	Fuyang	Dandong	Fuyang	Dandong	Fuyang	Dandong
Tfa01	$(AAAG)_{11}$	F: CAGCCTTACGGTGCAACATG R: AAAGTTGACCCAGAGGCGTC	202–318	18	17	0.9335	0.9477	0.9082	0.9233	0.9167	0.9583
Tfa02	$(AC)_{11}$	F: GTGCTGATACCAAAGGGCCT R: GCCATCCCAATAGTGGTCCC	207–219	5	5	0.4796	0.7119	0.4404	0.6494	0.4583	0.6250
Tfa06	$(AC)_{11}$	F: TGCAAAGGGACAGTACTGCA R: AGTAGCCGTACAGAGACGCA	215–239	7	8	0.6702	0.7633	0.6251	0.7115	0.6667	0.7917
Tfa08	$(AC)_{11}$	F: TCTCTAACAGTTGCGTCTCCT R: AACCCTCCACTGTCCAAACA	215–227	7	6	0.8333	0.8156	0.7904	0.7686	0.9167	0.7500
Tfa09	$(AC)_{11}$	F: AGTTAGTGCTTTGTTCCATTAAATGT R: CGGGTCGACTTGGTGATCAT	198–220	3	10	0.3511	0.8697	0.3067	0.8339	0.2083	0.9167
Tfa10	$(AC)_{11}$	F: TCAGGTCATGTTAGCGTGCA R: TGCTGCAGATGTCTCAGTGG	197–213	5	6	0.4158	0.7066	0.3728	0.6337	0.4167	0.7500
Tfa11	$(AG)_{11}$	F: CCTCCCTTCAGAAGCAGGTC R: GCTCGCTCTCCTCAATACCC	271–295	5	10	0.7562	0.8005	0.7011	0.7627	0.7500	0.7917
Tfa12	$(AC)_{11}$	F: GCTCCGGTGTCATATGCAGA R: TGTTTCTGCCGAATCCCACT	284–312	8	12	0.6782	0.8734	0.6135	0.8393	0.7083	0.7826
Tfa13	$(AC)_{11}$	F: ACAACTGGAGTGATGTCGGC R: CCTTCTGCAGTCCCTGTGTT	217–237	6	9	0.6676	0.8475	0.6098	0.8080	0.5833	0.8750
Tfa14	$(AC)_{11}$	F: CGTCTATCACTCATCGCAGACA R: GCCCATAATGGCGTTTGTTCT	127–185	7	14	0.7926	0.9229	0.7406	0.8957	0.7917	0.7500
Tfa22	$(AC)_{12}$	F: GACCGATGACCAGGTTACGG R: GAGATCAACGTGGTGGCTCA	268–304	7	11	0.6764	0.8870	0.6349	0.8540	0.6250	0.9565
Tfa23	$(AC)_{12}$	F: ACTATCACTACCCGTCTTTCCTC R: CGCTTTGATGCCATACTGCA	213–237	9	10	0.8077	0.8121	0.7637	0.7691	0.7391	0.7500
Tfa24	$(AAAG)_{12}$	F: GCACGCGTTTCCTCTTGTTT R: AGCACCTCACTGAGAATCGC	183–243	11	14	0.8954	0.8901	0.8651	0.8609	0.7083	0.8750
Tfa26	$(AC)_{12}$	F: TCTGGCAGAAAGGGCATGAA R: AGCAGAGAAGGTTAAGGCACC	193–221	8	9	0.8333	0.7793	0.7921	0.7288	0.7500	0.6667
Tfa27	$(AC)_{12}$	F: TCCAGCAGAGGATGTGTTTAGT R: CCGGACTTTGGCGTTGATTA	232–246	6	6	0.6613	0.7604	0.5835	0.7040	0.7500	0.7826
Tfa28	$(AC)_{12}$	F: GCAGCTTCAAGGCGATAGGA R: GAGCTCCTCATTACGCCCAA	165–199	9	13	0.8254	0.8932	0.7841	0.8628	0.7083	0.1364*
Tfa29	$(AC)_{12}$	F: ATTGGAGCAGGTCACCGTAG R: AGTGTCAGACCCAGATGTCCT	124–172	8	15	0.6605	0.9229	0.6224	0.8959	0.7083	0.9583
Tfa30	$(AC)_{12}$	F: CCATGCACCATTAGCCCTGT R: GTGAGTGACGTCGAAAGCGA	302–310	4	5	0.4716	0.6268	0.4247	0.5810	0.3333	0.5417
Tfa31	$(AC)_{12}$	F: GACTGTCTGCTGTCGACCAA R: CTGGTTCCCAGTAAGCTCCG	154–172	5	7	0.5993	0.7110	0.5501	0.6465	0.6667	0.6667

Table 2 continued

Locus	Repeat motif	Primer sequence (5′-3′)	Size range (bp)	*Na*		H_E		H_O		*PIC*	
				Fuyang[a]	Dandong[b]	Fuyang	Dandong	Fuyang	Dandong	Fuyang	Dandong
Tfa32	$(AC)_{12}$	F: TGATTGTCATTCCGCCGCTA R: TCTGCTCCAGCTCACTCTGA	234–272	7	9	0.8493	0.5230	0.8089	0.4976	0.6667*	0.2500*
Tfa33	$(AC)_{13}$	F: GTCCTTCCTGCCGAATGTGA R: AACGTCCGACCTCATCAGTG	143–193	8	13	0.8457	0.8298	0.8065	0.7911	0.8750	1.0000
Tfa35	$(AC)_{13}$	F: AGAACTCCCAGCCCAGGTT R: AGCCTGAACTTTGTCACACCT	138–214	9	19	0.7807	0.9291	0.7385	0.9031	0.8696	0.8333
Tfa36	$(AC)_{13}$	F: CCCAGTGCTGAGATTACCGG R: TCAGAGGGTGTTGTGCTGTG	232–274	9	5	0.8541	0.6853	0.8146	0.6120	0.6087*	0.7917
Tfa37	$(AC)_{13}$	F: CCTCAAATTACCTTGTGCTGTATCC R: GCCGGTCCTGGCCTTAAC	218–238	7	9	0.8191	0.7988	0.7728	0.7528	0.8333	0.7083
Tfa38	$(AC)_{13}$	F: GCATACACGGAAGTTGTCGAC R: CCACAGGAAGCAGGACTCAA	206–224	5	9	0.6693	0.8156	0.5936	0.7754	0.6667	0.7917
Tfa39	$(AAT)_{13}$	F: AAAGGGCCATTGTCGGAAGT R: AGTCTCAATGGATGCAAACACT	177–213	11	10	0.8696	0.8768	0.8341	0.8424	0.8696	0.8333
Tfa40	$(AC)_{13}$	F: GCCACTGCAGCTTTATTGCC R: AGACTGAGGTAGGGTCAGGG	146–178	8	13	0.6489	0.9016	0.5925	0.8720	0.6667	0.8750
Tfa41	$(AGG)_{13}$	F: CACAAGTTGGCTGGAGGTGA R: GATGGAGGCGATTACCCACC	226–259	9	6	0.8097	0.7287	0.7654	0.6598	0.5000	0.4583
Tfa42	$(AC)_{13}$	F: TGGGTTGGCTTTCACCTGAA R: CACCAACAAGACAGCTGTGG	217–307	12	21	0.8546	0.9193	0.8192	0.8934	0.8333	0.9167
Tfa43	$(AC)_{13}$	F: GACACGCTCTTCTGTCTGCT R: GTAGGCGTCCATGACAGGTC	233–259	10	7	0.5878	0.6628	0.5487	0.6035	0.3750	0.7391
Tfa44	$(AC)_{14}$	F: TCTCTTGCATGGACTGAACG R: CCCGGGCATTTCTCACAGAA	151–175	6	10	0.7402	0.8528	0.6749	0.8153	0.7500	0.7083
Tfa45	$(AC)_{14}$	F: TCTGCATCCCACTGTCAACG R: GGACGTTGAACATAGGCCCA	208–230	5	5	0.7562	0.7266	0.6962	0.6564	0.6250	0.6522
Tfa48	$(AC)_{14}$	F: GCACACCCACTTGTTTCCTG R: AGTGAATGAGTGCACACGCT	207–257	8	13	0.8147	0.8351	0.7698	0.7984	0.6250	0.7917
Tfa49	$(AC)_{15}$	F: TGCCTCCTCCAAGTTCACAC R: CCTGAGTCTGTCTAAAGCAACAC	278–302	7	9	0.6348	0.8280	0.5946	0.7871	0.5417	0.7500
Tfa50	$(AC)_{15}$	F: GAAGGTCTGACTCACCCACG R: AGATTATGTTTCTGCTCAAACTGGG	212–234	6	7	0.7926	0.6844	0.7411	0.6362	0.8750	0.6250
Tfa51	$(AG)_{15}$	F: GGCATCGTCTCATGACTGGT R: TCTCCTCGCTGTGTCAATGG	157–215	10	15	0.8520	0.8298	0.8135	0.7977	0.6667	0.6250
Tfa52	$(AC)_{15}$	F: ACCCGACGTTAGTGATGGTG R: TGCTTCACGTCCACTTCCTC	128–190	16	18	0.9069	0.9468	0.8797	0.9224	0.8750	0.9583
Tfa53	$(AC)_{15}$	F: GGTAATTACAGTGCAGCGGC R: CTTCTGCATGCGCTCCATTC	236–302	12	13	0.8670	0.9122	0.8332	0.8835	0.7083	0.8750
Tfa54	$(AC)_{16}$	F: GCGTCAGCAGGTATTGTCAC R: GCTGGACGGTTAGGGATGAC	245–311	12	19	0.8676	0.9353	0.8336	0.9099	0.5652	0.8750

Table 2 continued

Locus	Repeat motif	Primer sequence (5′-3′)	Size range (bp)	*Na*		H_E		H_O		*PIC*	
				Fuyang[a]	Dandong[b]	Fuyang	Dandong	Fuyang	Dandong	Fuyang	Dandong
Tfa55	$(AG)_{16}$	F: TGCGTATTACAGCTCCTGGC R: CCTTGGCTTTAGGTGGACGT	221–243	7	11	0.8209	0.8759	0.7746	0.8423	0.8333	0.9167
Tfa56	$(AAG)_{17}$	F: ACAGTTTCAACCTGTTCATGATCT R: CAGTAGATTTGAGTCCTCTTGGT	255–420	21	29	0.9527	0.9805	0.9279	0.9586	0.9130	0.9167
Tfa57	$(AAG)_{21}$	F: CCCAGCTGCTTCTTTCTCGT R: CGCTCGCCATATGCTCGATA	248–464	20	27	0.9433	0.9808	0.9187	0.9545	0.5000*	0.7000*
Tfa58	$(AAAC)_{7}$	F: GAGTTGCAAGTTCGAGTGGC R: ATCGTCAGTGTCAAACCGCT	224–236	3	4	0.5293	0.5470	0.4038	0.4765	0.5417	0.5000
Tfa59	$(AAAG)_{13}$	F: GCGTTAATGAATAACTCCTGAGCC R: TTCTCAGGAAACAGCCCAGA	244–344	16	20	0.9300	0.9477	0.9039	0.9234	1.0000	0.9167
Tfa62	$(AGAT)_{22}$	F: CGATGGTGGAAAGTTTGTGCC R: GGTCTGTTGCTTCATGACTGC	130–302	18	23	0.8183	0.9672	0.7919	0.9445	0.8333	0.8750

Ta annealing temperature of each primer pair, *Na* number of alleles observed, H_O observed heterozygosity, H_E expected heterozygosity, *PIC* polymorphism information content

* Significant deviation from HWE after Bonferroni correction ($P < 0.0011$)

[a] Fuyang, Zhejiang Province, China (30°03′N, 119°58′E)

[b] Dandong, Liaoning Province, China (39°46′N, 124°20′E)

and cost-effective microsatellite identification approach (RAD-seq-Assembly-Microsatellite) using paired-end RAD-seq. This approach can create longer sequences by assembling the overlapping paired-end RAD reads for microsatellite discovery, which overcomes the issues of short read lengths generated by the Illumina platform and improves microsatellite detection rates. The approach could efficiently generate a large set of polymorphic microsatellite markers for a wide range of applications from population genetics, behavioral ecology, to marker-based breeding programs, especially for non-model species.

Next-generation sequencing (NGS) technologies have enhanced our ability to obtain hundreds of microsatellites in a rapid and low cost manner, and dramatically accelerated the discovery of genomic information even in non-model species (Cai et al. 2013; Davey et al. 2011; Hu et al. 2016). Compared with traditional methods for microsatellite markers development (Zane et al. 2002), our approach, which based on RAD-seq and de novo local assembly, is more efficient in terms of money and time. At current market prices, the RAD library construction costs approximately $75, and a 125 bp × 2 paired-end sequencing run on the Illumina Hiseq 2500 platform to produce 1 Gb of sequences costs approximately $70. The Illumina Hiseq 4000 and X-ten platforms have much lower per-base sequencing costs with higher throughout than Hiseq 2500 platform (http://www.illumina.com). So the total costs for library construction and sequencing using our approach ($360 for 4 Gb data) are lower than that using traditional (at least $800 for 100 sequence data) approaches (Zalapa et al. 2012). In general, traditional approaches require 2–4 weeks for DNA extraction, library construction, cloning, Sanger sequencing and primer design, whereas our approach requires only 1–2 weeks for DNA extraction, RAD library construction, Illumina sequencing and primer design. Moreover, our approach can identify thousands of microsatellites and then batch design primers simultaneously, while traditional approaches can usually find repeated units in a relative small pool of sequences (Zane et al. 2002).

There have been studies that used the next generation sequencing technologies to develop microsatellite markers (Bonatelli et al. 2015; Castoe et al. 2010; Hung et al. 2016; Li et al. 2016a, b; Minegishi et al. 2015). The two major NGS platforms used for the discovery of microsatellites are Roche 454 and Illumina sequencing (Zalapa et al. 2012). The main advantage of the 454 sequencing for microsatellite discovery is that the read-length is 350–600 bp, which could allow the discovery and development of microsatellites even directly from the raw reads. Castoe et al. (2010) identified 14,612 microsatellite loci in 11.3% of the 128,773 Roche 454 shotgun reads, and

4564 of which had flanking sequences suitable for primer design. Bonatelli et al. (2015) validated 22 (30.56%) polymorphic microsatellites out of 64 loci using double digest restriction site–associated DNA sequencing (ddRAD-seq) on a Roche 454 platform. Seventy-four projects using 454 sequencing reviewed in Hodel et al. (2016) yielded 8–91 polymorphic loci, with an average of 16 polymorphic loci and 4400 potential loci derived from an average of 139,418 reads. The main advantage of using Illumina platform for microsatellite discovery is that the much higher throughout and lower costs than 454 platform (Zalapa et al. 2012). Cai et al. (2013) assembled the generated Illumina shotgun reads into a draft genome of *Anisogramma anomala*, and successfully amplified 214 (90.7%) microsatellite loci with specific products. Hung et al. (2016) applied Illumina shotgun sequencing to *Apodemus semotus* and mapped the obtained sequences reads against the genome of *Mus musulus*, then successfully amplified 44 (74.57%) of 59 microsatellite loci. Hu et al. (2016) used transcriptome data from RNA-Seq using Illumina sequencing and found that 20 (31.75%) loci were successfully amplified and also were polymormic. For microsatellites development studies using Illumina sequencing reviewed in Hodel et al. (2016), the average number of polymorphic microsatellite markers reported was 15, and the average number of potential loci per study was 15,539.

However, the Illumina platform generates relative short reads (100–300 bp), so assembly is usually required to achieve longer contiguous sequences, which could provide sufficient flanking sequence for the design of primers to amplify the target microsatellite and reduce redundancy of closely linked microsatellites (Zalapa et al. 2012). Yang et al. (2016) only identified 650 microsatellite loci from 4.5 million RAD raw reads and only 285 (43.84%) primer pairs were successfully designed. However, in the present study, 22,835 microsatellites were discovered in 121,750 contigs assembled, and 156,150 primer pairs were designed for 16,497 (77.24%) loci containing microsatellites. Therefore, careful consideration should be given to the quality of assembly. RAD-seq is a family of genomic approaches that provide sequence data adjacent to restriction enzyme recognition sites (Davey et al. 2011; Hohenlohe et al. 2013). Furthermore, the overlapping paired-end reads by traditional RAD-seq technology allowed local assembly of contigs containing both the forward and reverse reads of each pair. These RAD contigs are anchored at one end by the restriction enzyme recognition site and contain several hundred base pairs of continuous genomic sequence data (Hohenlohe et al. 2013). This assembly method for RAD holds several advantages comparing to whole genome sequencing. Firstly, reads of a single RAD locus could be clustered before assembly using the similarity of the first reads with the restriction enzyme recognition site, and therefore reducing complexity and the computational costs, which is even affordable for desktop computer. Secondly, local assembly of the contigs could improve the quality of de novo assembly, and therefore improve the success rate of microsatellite development. Thirdly, the size selection in library preparation is flexible, which makes the length of contigs easily customized. In our study, PCR amplifications were successful for 48 (92.31%) of the 52 randomly selected loci, which was higher than those of the other approaches above. This indicated that primer pairs in the database and the assembled RAD contigs were of high quality and most of primer pairs would amplify their targets. Compared to other approaches using next-generation sequencing, our assembly based approach exhibited great advantages on developing thousands of microsatellites rapidly and accurately, especially for non-model species with shallow background of genomic information.

In conclusion, the present study has contributed a detailed approach to rapidly and cost-effectively develop genome-wide microsatellite markers in non-model species with high success rate. A total of 45 polymorphic loci were validated, which could serve as a proof-of-concept showing that the "RAD-seq-Assembly-Microsatellite" approach was successfully applied to a non-model species. The "RAD-seq-Assembly-Microsatellite" approach developed in the present study holds great promise for microsatellite development in future ecological and evolutionary studies of non-model species.

Acknowledgements We are thankful to the anonymous referee for critical and helpful comments on the manuscript.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval All applicable international, national, and/or institutional guidelines for the care and use of animals were followed.

Funding This study was funded by the National Natural Science Foundation of China (NSFC) (Grant number 31502169), AoShan Talents Program supported by Qingdao National Laboratory of Marine Science and Technology (Grant Number 2015ASTP-ES05), the NSFC-Shandong Joint Fund for Marine Ecology and Environmental Sciences (Grant Number U1606404), and AoShan Science and Technology Innovation Plan (Grant Number 2015ASKJ02-05).

References

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche

E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3–W10

Bonatelli IA, Carstens BC, Moraes EM (2015) Using next generation RAD sequencing to isolate multispecies microsatellites for *Pilosocereus* (Cactaceae). PLoS One 10:e0142602

Cai G, Leadbetter CW, Muehlbauer MF, Molnar TJ, Hillman BI (2013) Genome-wide microsatellite identification in the fungus *Anisogramma anomala* using Illumina sequencing and genome assembly. PLoS One 8:e82408

Castoe TA, Poole AW, Gu W, Jason de Koning AP, Daza JM, Smith EN, Pollock DD (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Mol Ecol Resour 10:341–347

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140

Chang Y, Feng Z, Yu J, Ding J (2009) Genetic variability analysis in five populations of the sea cucumber *Stichopus (Apostichopus) japonicus* from China, Russia, South Korea and Japan as revealed by microsatellite markers. Mar Ecol 30:455–461

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759–769

Glenn TC, Schable NA (2005) Isolating microsatellite DNA loci. Methods Enzymol 395:202–222

Hodel RG, Segovia-Salcedo MC, Landis JB, Crowl AA, Sun M, Liu X, Gitzendanner MA, Douglas NA, Germain-Aubrey CC, Chen S, Soltis DE, Soltis PS (2016) The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. Appl Plant Sci 4:1600025

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol Ecol 22:3002–3013

Hu Z, Zhang T, Gao X-X, Wang Y, Zhang Q, Zhou H-J, Zhao G-F, Wang M-L, Woeste KE, Zhao P (2016) De novo assembly and characterization of the leaf, bud, and fruit transcriptome from the vulnerable tree *Juglans mandshurica* for the development of 20 new microsatellite markers using Illumina sequencing. Mol Genet Genom 291:849–862

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Hung C-M, Yu A-Y, Lai Y-T, Shaner P-JL (2016) Developing informative microsatellite makers for non-model species using reference mapping against a model species' genome. Sci Rep 6:23087

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303:3997 (**Prepr**)

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li Q-Y, Zhang J, Yao J-T, Wang X-L, Duan D-L (2016a) Development of *Saccharina japonica* genomic SSR markers using next-generation sequencing. J Appl Phycol 28:1387–1390

Li Y-L, Xue D-X, Gao T-X, Liu J-X (2016b) Genetic diversity and population structure of the roughskin sculpin (*Trachidermus fasciatus* Heckel) inferred from microsatellite analyses: implications for its conservation and management. Conserv Genet 17:921–930

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol Ecol Resour 15:28–41

Meglécz E, Costedoat C, Dubut V, Gilles A, Malausa T, Pech N, Martin JF (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. Bioinformatics 26:403–404

Minegishi Y, Ikeda M, Kijima A (2015) Novel microsatellite marker development from the unassembled genome sequence data of the marbled flounder *Pseudopleuronectes yokohamae*. Mar Genom 24:357–361

Montanari S, Perchepied L, Renault D, Frijters L, Velasco R, Horner M, Gardiner SE, Chagné D, Bus VGM, Durel C-E, Malnoy M (2016) A QTL detected in an interspecific pear population confers stable fire blight resistance across different environments and genetic backgrounds. Mol Breed 36:1–16

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 29:294–307

Onikura N, Takeshita N, Matsui S, Kimura S (2002) Spawning grounds and nests of *Trachidermus fasciatus* (Cottidae) in the Kashima and Shiota estuaries system facing Ariake Bay, Japan. Ichthyol Res 49:198–201

Park S (2001) Trypanotolerance in West African cattle and the population genetic effects of selection. Dissertation, University of Dublin

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8:103–106

Schoebel CN, Brodbeck S, Buehler D, Cornejo C, Gajurel J, Hartikainen H, Keller D, Leys M, Říčanová S, Segelbacher G, Werth S, Csencsics D (2013) Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. J Evol Biol 26:600–611

Stabile J, Lipus D, Maceda L, Maltz M, Roy N, Wirgin I (2016) Microsatellite DNA analysis of spatial and temporal population structuring of *Phragmites australis* along the Hudson River Estuary. Biol Invasions 18:2517–2519

van Oosterhout C, Hutchinson WF, Wills DP, Shipley P (2004) Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538

Wang J-Q (1999) Advances in studies on the ecology and reproductive biology of *Trachidermus fasciatus* Heckel. Acta Hydrobiol Sin 23:729–734 **(in Chinese)**

Wang J-Q, Cheng G (2010) The historical variance and causes of geographical distribution of a roughskin sculpin (*Trachidermus fasciatus* Heckel) in Chinese territory. Acta Ecol Sin 30:6845–6853 **(in Chinese)**

Weinman LR, Solomon JW, Rubenstein DR (2015) A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. Mol Ecol Resour 15:502–511

Xu J-R, Han X-L, Li N, Yu J-F, Xu P, Bao Z-M (2008) Analysis of genetic diversity in roughskin sculpin *Trachidermus fasciatus* by AFLP markers. J Dalian Fish Univ 23:437–441 **(in Chinese)**

Xue D-X, Zhang T, Liu J-X (2014) Microsatellite evidence for high frequency of multiple paternity in the marine gastropod *Rapana venosa*. PLoS One 9:e86508

Yang X-Y, Long Z-C, Gichira AW, Guo Y-H, Wang Q-F, Chen J-M (2016) Development of microsatellite markers in the tetraploid fern *Ceratopteris thalictroides* (Parkeriaceae) using RAD tag sequencing. Genet Mol Res. doi:10.4238/gmr.15017550

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11:1–16

Zeng Z, Liu Z-Z, Pan L-D, Tang W-Q, Wang Q, Geng Y-H (2012) Analysis of genetic diversity in wild populations of *Trachidermus fasciatus* by RAPD and the transformation of two SCAR markers. Zool Res 33:203–210