

# Distribution, Function and Polymorphism Characteristics of Microsatellites in *Pyropia yezoensis* Transcriptome

LIU Yang<sup>1), 3)</sup>, PAN Xue<sup>1), 3)</sup>, XU Kuipeng<sup>1), 3)</sup>, and MAO Yunxiang<sup>1), 2), 3), \*</sup>

1) Key Laboratory of Marine Genetics and Breeding (Ocean University of China), Ministry of Education, Qingdao 266003, China

2) Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

3) College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China

(Received March 9, 2018; revised May 7, 2018; accepted December 16, 2018)

© Ocean University of China, Science Press and Springer-Verlag GmbH Germany 2019

**Abstract** The distribution, putative function and polymorphism characteristics of simple sequence repeats (SSRs) in *P. yezoensis* transcriptome were analyzed in this study. In total, 3076 SSRs were detected among 2681 unigenes. Trinucleotide types were dominant, constituting 91.87% of all the microsatellites. The most abundant SSR was CCG (71.03%) and the second abundant one was AGC (234, 7.61%). A total of 111 (3.61%) dinucleotide types were found and the most abundant one was AC (51, 1.66%) which was followed by CG (34, 1.11%). SSRs identified showed a GC bases preference; GC bases constituted 89.73% of all the SSR bases. SSRs occurrence number decreased as repetitions increased. Annotation results exhibited that the majority of SSRs-containing unigenes have the functions of 'metabolic process', 'binding' and 'catalytic activity' and existed as the forms of 'cell', 'cell component' and 'organelle'. The dominant amino acids that SSRs coded were Ala (28.59%), Arg (26.02%), Gly (14.35%) and Pro (13.65%). Amplification results showed that 3 out 23 SSRs from transcriptome (13.04%) and 8 out 77 SSRs from genome (10.39%) were polymorphic.

**Key words** *P. yezoensis*; transcriptome; SSR; distribution; function; polymorphism

## 1 Introduction

Microsatellites or simple sequence repeats (SSRs) are clusters of short tandem repeated nucleotides distributed throughout the genome (Hans, 2004). More and more researches have been showing that SSRs are non-randomly distributed in the genome (Li *et al.*, 2004). Numerous SSRs have been found in both coding and non-coding regions (Tóth *et al.*, 2000). SSRs, especially those located in coding region, may play important roles in chromatin organization, gene expression regulation, DNA replication, cell cycle, species evolution, environmental stress adaptability and so on (Li *et al.*, 2004; Kashi and King, 2006).

SSRs are more variable than other DNA sequence types in genome. SSR polymorphism is considered as the result of strand-slippage during DNA replication process (Katti *et al.*, 2001), deriving mainly from the variability in length (Hans, 2004). The component bases in SSRs tend to positively correlate with GC content of DNA sequences. In the coding regions of most plant genes, the most frequent SSRs type is AAG. However, in cereals the

most common type is CCG. It was found that CCG repeat is abundant in monocots coding regions, which may be due to the high GC content (Kalia *et al.*, 2011).

SSR markers have been widely used to construct the genetic linkage maps and analyze the genetic diversity as they are dominant, abundant, multi-allelic and easily detectable (Liu *et al.*, 2015). The SSR markers obtained from coding region represent even more useful information than other genomic DNA-based markers. These markers can be used to detect the variation in the expression region and can be used as trait-marker in MAS (Marker-Assisted Selection) (Victoria *et al.*, 2011).

*Pyropia yezoensis*, belonging to Rhodophyta, is an economically important marine crop grown in intertidal habitat. It widely spreads in the coasts of China, Japan and Korea. In its life cycle, *P. yezoensis* suffers from the stress of temperature change, osmotic pressure and ultraviolet radiation (Sun *et al.*, 2015). It has a biphasic life cycle, including gametophytic blades (haploid) and conchocelis phase (diploid). The conchocelis could be cultured in seawater or liquid culture medium, named as free-living conchocelis (Yan and Aruga, 2000). The *P. yezoensis* transcriptome and genome have been assembled (Yoji *et al.*, 2013; Sun *et al.*, 2015). However, little research has been conducted focusing on the SSR distri-

\* Corresponding author. Tel: 0086-532-82032017

E-mail: yxmao@ouc.edu.cn

bution and function characteristics in *P. yezoensis* transcriptome. Present research systemically analyzed the SSR distribution characteristics in *P. yezoensis* transcriptome, annotated the functions of unigenes containing SSRs and analyzed the amino acids characteristics coded by SSRs. To evaluate the polymorphism of SSRs, 77 pairs of primers were designed from *P. yezoensis* genome to analyze 12 free-living conchocelis strains cultured as the alga germplasm.

## 2 Materials and Method

### 2.1 Analysis of SSR Distribution Characteristics

A total of 18734 non-redundant *P. yezoensis* unigenes were obtained from our previous study (Sun *et al.*, 2015). Due to the effect of polyA tail, only SSRs constituted with 2 to 6 bases were analyzed using MISA tool (Hans, 2004). Furthermore, the dinucleotide SSR types should repeat at least 7 times, trinucleotide 5 times, tetranucleotide 4 times, pentanucleotide 4 times and hexanucleotide 4 times. The criterion for compound SSRs was that the interval length between two loci was shorter than 100bp. Parameter SSR density (No./Mbp) was employed to evaluate the abundance of SSRs. The data set of *P. yezoensis* genome was downloaded from website (<http://nrifs.fra.affrc.go.jp>) and analyzed using the same pipeline above.

### 2.2 Annotation of SSR-Containing Unigenes

Blastall software was used to align (E-value < 1E-5) the SSR-containing unigenes to Nr (Non-redundant) protein database, KEGG (Kyoto Encyclopedia of Genes and Genomes), Swiss-Prot and KOG (Eukaryotic Orthologous Groups) databases, respectively. The annotation results acquired from Nr were processed through Blast2GO program to obtain the relevant GO (Gene Ontology) terms, and then were analyzed by WEGO software to assign GO functions (Wang *et al.*, 2013). Annotation results revealed that the SSR-containing unigenes aligned to the Swiss-Prot database were the most abundant. These unigenes were further used to predict ORF regions and analyze the characteristics of amino acids coded by SSR.

### 2.3 Primer Designing and Validation for SSR Markers

Basing on the previous SSRs analysis results, 1962 pairs of primers were designed from *P. yezoensis* genome using Primer 3, and 77 pairs were randomly selected to test the availability in 12 free-living conchocelis strains cultured in our laboratory. The free-living strains were obtained from germplasm strains collected from MouPing (37°28.00'N, 121°37.53'E), marked as MZ and MK; Peng-Lai (37°49.93'N, 120°44.73'E), marked as PZ and PK; TuanDao (36°3.22'N, 120°17.55'E), marked as TZ and TK; MaTiJiao (36°4.28'N, 120°17.88'E), marked as JZ and JK; HuiQuanWan (36°2.73'N, 120°20.17'E), marked as HZ and HK; DaLian (38°52.58'N, 121°33.72'E), marked as DZ and DK.

DNA of free-living conchocelis strains were extracted using a plant genomic DNA extraction kit (Tiangen Bio-

tech Co., Ltd., Beijing, China) according to the manufacturer's instructions. Polymerase chain reaction (PCR) was carried out in a volume of 25  $\mu$ L (1.0 U *Taq* DNA polymerase, 2.5  $\mu$ L 10 $\times$ PCR buffer, 5 ng template DNA, 0.2  $\mu$ mol L<sup>-1</sup> each primer and 200  $\mu$ mol L<sup>-1</sup> dNTP). The mixtures were subjected to PCR machine with a procedure of 94°C for 10 min, followed by 35 cycles of 94°C for 30 s, annealing at temperatures appropriate for primer pairs each for 30 s, 72°C for 30 s and 72°C for 10 min. The PCR product was detected with 8.0% non-denaturing polyacrylamide gel and then visualized by silver-staining. The locus was considered as polymorphic if more than one band was detected at the same position for all the strains (Bi *et al.*, 2014).

## 3 Results

### 3.1 SSR Distribution Characteristics

Analysis showed that 3076 SSRs existed among 2681 unigenes, accounting for 14.31% of the total unigenes, and the SSR density was 217.53 per Mbp, about 4.60 kb each (Table 1). Data analysis revealed 68 SSRs types. Trinucleotide type was dominant, constituting 91.87% of the total. The most abundant repeat element was CCG (2185, 71.03%) and the second most abundant one was AGC (234, 7.61%) (Fig.1). A total of 111 (3.61%) dinucleotide types were found, while the most abundant repeat element was AC (51, 1.66%), followed by CG (34, 1.11%). Pentanucleotide types were the least abundant and only 9 (0.29%) were detected. In total, 74 hexanucleotide types were found (2.41%). SSRs showed a GC bases preference and GC bases constituted 89.73% of all the SSR bases.

Table 1 SSR distribution characteristics in *P. yezoensis* transcriptome

Item	<i>P. yezoensis</i>
Unigene number	18734
Size of transcriptome (bp)	14140619
GC content (%)	67.78
SSRs number	3076
SSRs density (No./Mbp)	217.53

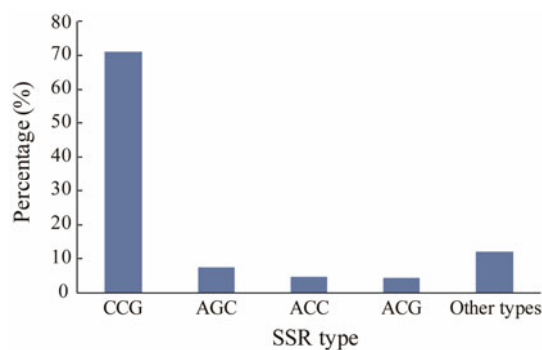


Fig.1 Percentage of SSR types in *P. yezoensis* transcriptome.

Analysis found that SSR occurrence number decreased as repetition increased (Fig.2). The only exception was that repetition 11 (11) dinucleotide loci occurred more than repetitions 9 (5) and 10 (2). The length of the most

abundant SSRs was 15 bp, followed by 18 bp (Fig.3). This phenomenon was due to the dominant position of trinucleotide types. The longest was (CA)<sub>50</sub> with a size of 100 bp.

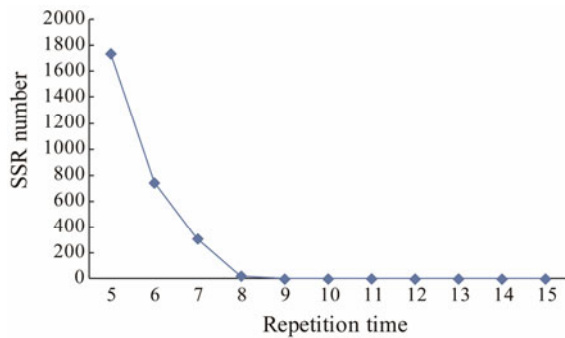


Fig.2 Variation of trinucleotide loci for different repetitions in *P. yezoensis* transcriptome.

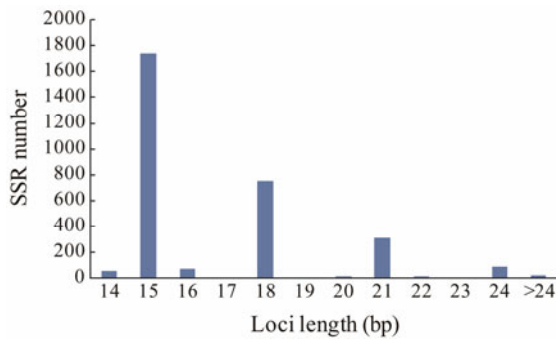


Fig.3 Length distribution of SSRs in *P. yezoensis* transcriptome.

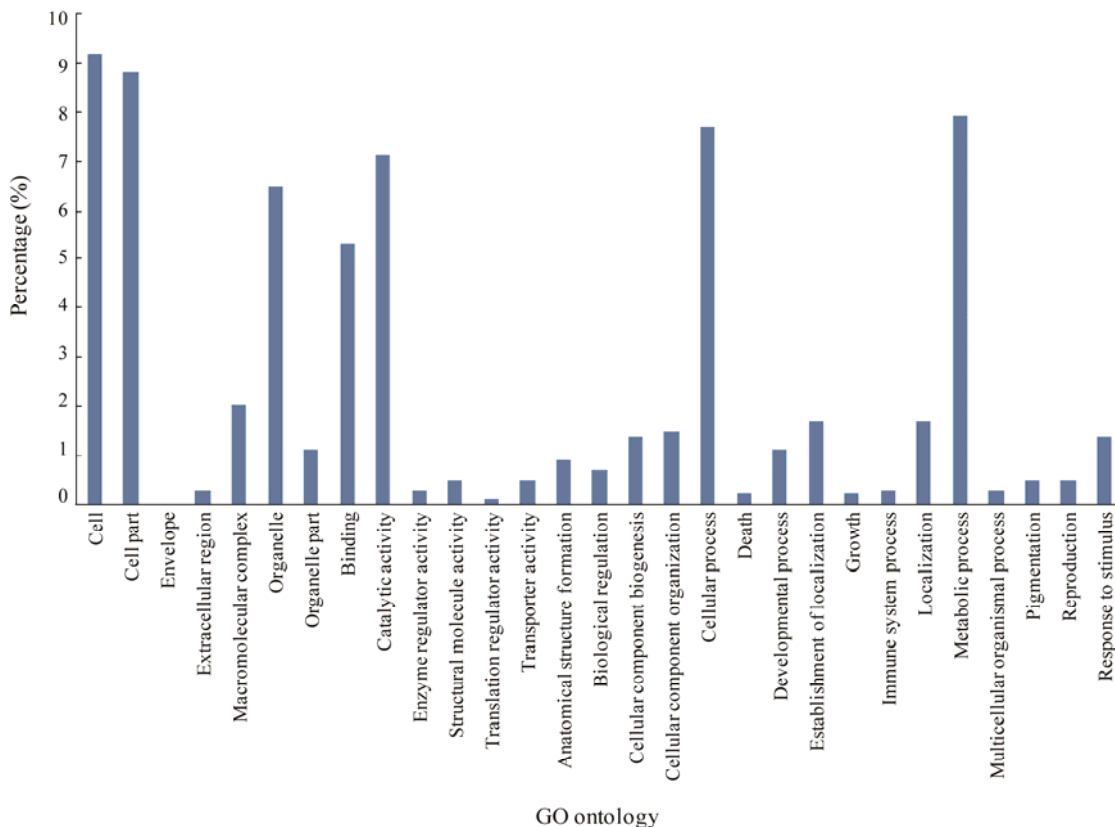


Fig.4 GO functional classification of SSR-containing unigenes in *P. yezoensis* transcriptome.

### 3.2 Annotation Results of SSR-Containing Unigenes

Unigenes containing SSR were aligned to the databases (GO, KEGG, KOG and Swiss-Prot) and their putative functions were characterized. GO annotation provides a system to categorize the gene products according to three ontologies: Cellular Component, Biological Process, and Molecular Function (Wang *et al.*, 2013). From the GO annotation, it was found that 545 (20.33%) SSR-containing unigenes matched to the Nr database. A total of 12.92% annotated unigenes were assigned to the ‘cell’ part of Cellular Component ontology; 12.45% to the ‘cell’ part and 9.09% to ‘cell organelle’ part (Fig.4). In the Biological Process ontology, ‘metabolic process’ (11.19%) and ‘cellular process’ (10.87%) were the dominant parts in SSR-containing unigenes. In the Molecular Function ontology, ‘catalytic activity’ (9.39%) and ‘binding’ (7.41%) were the dominant parts.

KEGG database records the networks of molecular interactions in the cell and variants of them specific to particular organisms. KEGG annotation showed that 520 (16.91%) SSR-containing unigenes could be assigned to the database, annotating to 204 pathways. A total of 17.73% of the annotated SSR-containing unigenes belonged to ‘signal transduction’. Other majority pathways were ‘carbohydrate metabolism’ (5.07%), ‘amino acid metabolism’ (4.46%), ‘translation’ (5.91%) and ‘folding storage and degradation’ (5.31%).

KOG annotation is used to predict and classify possible functions and 219 (7.12%) SSR-containing unigenes could be assigned to the KOG database, annotating 1040

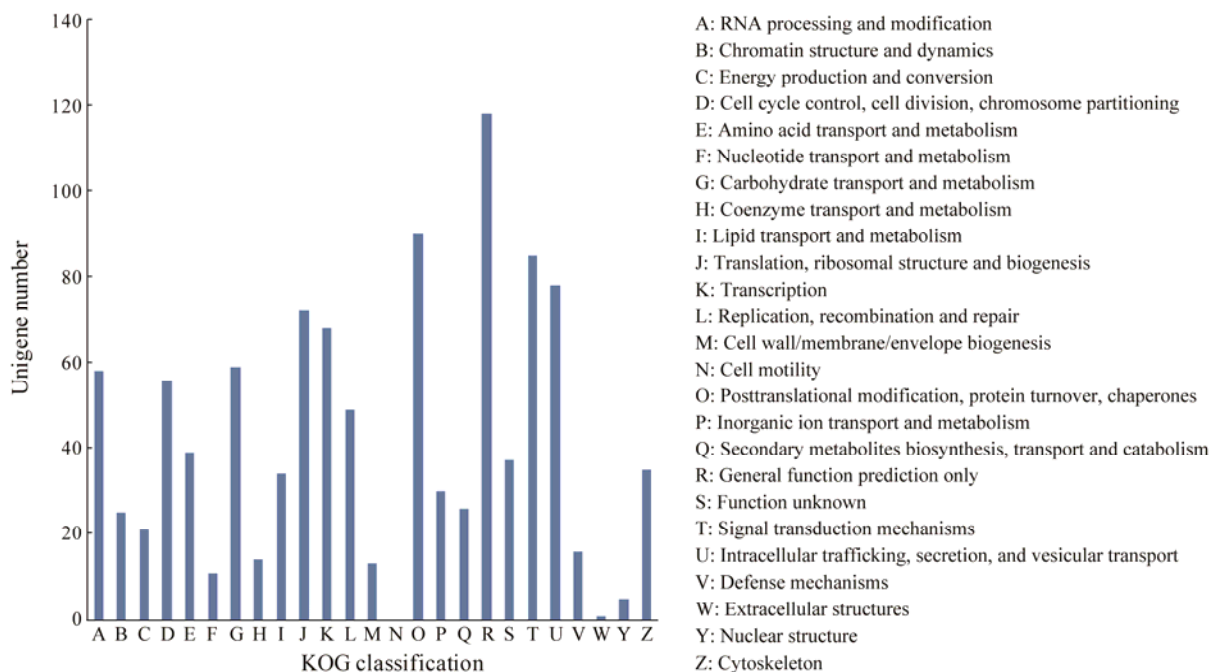


Fig.5 KOG functional classifications of SSR-containing unigenes in *P. yezoensis* transcriptome.

functions, respectively. Among 25 KOG classifications, the cluster 'General function prediction only' (11.35%) represented the largest group, and the following clusters were 'Posttranslational modification, protein turnover, chaperones' (8.65%), 'Intracellular trafficking, secretion, and vesicular transport' (7.50%), 'Signal transduction mechanisms' (8.17%), 'Translation, ribosomal structure and biogenesis' (6.92%) and 'Transcription' (6.54%) (Fig.5).

Swiss-Prot is the database that all the proteins have been annotated, and each protein has detailed sequence information. Annotation result showed that 750 (24.38%) SSRs-containing unigenes could be assigned to the Swiss-Prot database. And unigenes that aligned to *Arabidopsis thaliana* were the most abundant (26.67%), followed by *Dictyostelium discoideum* (7.33%), *Homo sapiens* (7.33%) and *Oryza sativa* (5.60%).

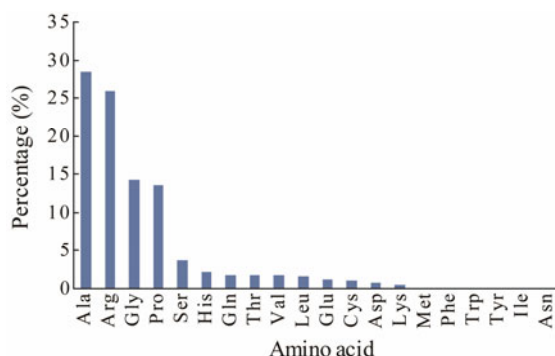


Fig.6 Amino acids coded by SSRs in *P. yezoensis* transcriptome.

Aforementioned, SSR-containing unigenes that assigned to the Swiss-Prot database were the most abundant. These unigenes were used to predict ORF and analyze the characteristics of amino acids coded by SSRs loci. The results showed that there were 857 SSRs coding amino

acids in these unigenes. Ala (28.59%), Arg (26.02%), Gly (14.35%) and Pro (13.65%) were the majority amino acids (Fig.6). The first two nucleotides in the codons of these four amino acids were G or C base. Thus this result was correlated with the preference of GC priority in *P. yezoensis* transcriptome.

### 3.3 Polymorphism Evaluation of SSR Markers

To evaluate the polymorphism of SSRs, 1962 pairs of primers were designed from *P. yezoensis* genome using the program of Primer 3. A total of 77 pairs were randomly selected to test the availability of the genes in 12 free-living conchocelis strains. Among the 77 pairs of primers, 23 pairs located in *P. yezoensis* transcriptome. The amplification results showed that for 36 (46.75%) pairs there was no amplification in all strains. For the other 41 (53.25%) pairs of amplified primers, with 20 pairs the same band could be amplified in all of the strains. How-

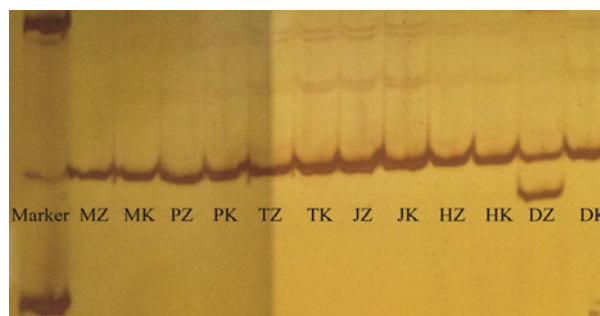


Fig.7 Amplification result of primer X3-28. MZ and MK: collected in MouPing (37°28.00'N, 121°37.53' E); PZ and PK: collected in PengLai (37°49.93'N, 120°44.73' E); TZ and TK: collected in TuanDao (36°3.22'N, 120°17.55' E); JZ and JK: collected in MaTiJiao (36°4.28'N, 120°17.88' E); HZ and HK: collected in HuiQuanWan (36°2.73'N, 120°20.17' E); DZ and DK: collected in DaLian (38°52.58'N, 121°33.72' E).

ever, amplification results of DZ strain showed that 6 pairs (X3-19, X3-25, X3-28, L9, L32 and L46) detected two bands, respectively (Fig.7), which means that DZ strain was heterozygous conchocelis. In other 11 strains only one band could be detected by one pair of primers, exhibiting purity conchocelis. The locus was considered as polymorphism if more than one band was detected in all strains (Bi *et al.*, 2014). Including the heterozygous

DZ strain, 8 (10.39%) (X3-19, X3-25, X3-26, X3-27, X3-28, L9, L32 and L46) out 77 pairs exhibited polymorphic in detecting 13 strains (considering heterozygous DZ as two purity strains). With 23 pairs of primers located in transcriptome, only 3 (13.04%) pairs (X3-27, X3-28 and L32) exhibited polymorphic.

The SSRs that could be detected by primers were related with the unigenes annotation results (Table 2).

Table 2 Primer sequences generated for SSRs amplification in *P. yezoensis*

Locus	Repeat motif	Primer sequence (5'-3')	Product size (bp)	Unigene function
L16	(TGG)5	F AGTCTCGTCGACGCTACCTC R ACATCATCCTCAGGAGGCAC	199	NA
L30	(CTC)7	F CTGGTTAAAGCCAACGTGGT R CCAGCGGTGTCGTACATAGA	237	Histidinol dehydrogenase
L32	(CGTG)4	F AGACGGACTTTTGAGGAGGG R GGCTGTACAACAGGAATGGG	153	Photosystem II M
X4-3	(GCTG)4	F GACTGTATACGCGCCTGTGA R AGCGTCAAGGGTATAGGCAA	226	NA
X4-4	(AAC)5	F AACACGCAAGCACACATTTC R TACTGGATGTGGAAGGGGAG	255	Protochlorophyllide oxidoreductase
X3-1	(GGT)5	F GAAGTGGCGTAATGCTGTTC R CTACCTCCTCTCAACCACGC	174	NA
X3-2	(AGA)5	F CCTTAAGAAGCCAGCGAGAA R TCTGAGGACGAGAGCAGTGA	223	Eukaryotic translation initiation factor 3 subunit C
X3-6	(CTG)5	F GACGGGAGAAAGACGAGTTG R GTAGGAATATGGCCGCAAAG	111	Unnamed protein product
X3-7	(GCT)6	F CCCCTACCAGGAAAGGAGAC R TTTCATGAAGATGACGACGC	211	Cytochrome c oxidase assembly COX19
X3-10	(GTC)5	F CCAGGCCCTTCATCAGGAGTA R ACCACCACTCTCACTACGGC	230	NA
X3-14	(CAG)5	F CACAACCTGGAGCACCTCAA R CCGACATGTATTGCGAGCTA	255	Importin alpha
X3-23	(TGC)6	F TTGCTGGTAGGATCCCAAG R GAACTGACGAGCAAAGAGGC	191	NA
X3-24	(CCG)5	F GACTGGAGGGTATTGCCAG R ATGGTAGTGGTAGCGGCATT	270	Principal sigma factor
X3-27	(GTG)5	F GCGACGGATAAGAAGAGTGC R TCGTTCATGATAGCCACTGC	152	NA
X3-28	(CCG)5	F CCAAGTCTTCGAGTCGTCC R GGGAGCATGGAGTACCTGTC	274	Alpha-galactosidase
X3-29	(GGC)6	F GCTCCTACCTCGCCCTCTAT R GATCACGTCGAGGATGTTCA	116	RNA polymerase sigma factor
X3-30	(TCG)5	F AGCGGGTCAGATCTTTACGA R CGAAGAAGTGAAGACCCTG	132	Kinase domain containing
X3-31	(GGC)6	F CAGACTCATGCCAACGAAGA R TATGCATCATTGTCCCCAGA	194	Vacuolar sorting 13
X3-32	(CCGC)4	F TCCGTGTGGATGACAATCTC R CAGAACTGTTGAGGAGGGC	190	Methyltransferase type 11

Note: NA, not available.

## 4 Discussion

Many researches have suggested that SSRs, especially those located in coding region, may play important roles in regulating gene expression (Li *et al.*, 2004). The mutation of SSRs located in the coding region of *P. yezoensis* may help the species to adapt the diverse intertidal environment in a fast way. This research firstly revealed distribution characteristics, putative function and evaluated the polymorphism of SSRs in *P. yezoensis* transcriptome.

In many plant transcriptomes, unigenes contained more trinucleotide and hexanucleotide SSRs, than dinucleotide and tetranucleotide. The frequency analysis of SSR types in *P. yezoensis* transcriptome found that trinucleotide types constituted 91.87% of all the loci. *P. haitanensis*, *P. seriata* and *P. tenera* were species of genus *Pyropia*. Although the criteria for SSRs screening in different research varied, trinucleotide types were the dominant loci in three ahead species transcriptomes, accounting for 87.17%, 94.10% and 90.20%, respectively (Xie *et al.*, 2013; Choi *et al.*, 2013; Im *et al.*, 2015). The dominance of tri-

nucleotide over other SSR types in transcribed regions may be because the trinucleotide can less likely to cause frame shift mutations (Li *et al.*, 2004; Liu *et al.*, 2015). Dinucleotide was the second most abundant SSR type in *P. yezoensis* transcriptome. The dinucleotide loci occurred most were AC. In *Sargassum thunbergii* transcriptome and *Chlamydomonas reinhardtii* EST (Expressed Sequence Tags) sequences, AC was also the abundant dinucleotide types (Stackelberg *et al.*, 2006; Liu *et al.*, 2015).

Analysis from some species transcriptomes found that SSRs in transcribed region showed a preference to special nucleotide. In *A. thaliana* and *Brassica rapa*, the most abundant locus was AAG (Biswas *et al.*, 2004; Asadi and Monfared, 2014). In cereal species, wheat and sorghum, the most common loci was CCG (Reddy *et al.*, 2012). The results showed that CCG repeat was abundant in monocots and it may be due to the high GC content (Kalia *et al.*, 2011). GGC was the predominant in the rice genome (Mun *et al.*, 2006). Researcher deemed that the widespread occurrence of GC-rich trinucleotides (those which contain  $\geq 2$ G and/or C in their repeating units) in expressed sequences of rice and maize reflected the overall high GC content of coding region in species from *Gramineae* family (Temnykh *et al.*, 2001). Some algae and Bryophytes species also exhibited the correlation between SSR type preference and GC content. The most abundant trinucleotide type in EST sequences of *C. reinhardtii* and *Physcomitrella patens* was CCG and reflected the high GC content in these two species (Victoria *et al.*, 2011). In *P. yezoensis* transcriptome, CCG constituted 71.03% of all the SSRs loci. This phenomenon was similar with that in *P. haitanensis* transcriptome, CCG accounting for 68.91% (Xie *et al.*, 2013). In *P. seriata* and *P. tenera* transcriptomes, the most common loci were GGC (Choi *et al.*, 2013; Im *et al.*, 2015). The transcriptomes of four species in genus *Pyropia* exhibited high GC content pattern (*P. yezoensis* 67.78, *P. haitanensis* 63.99, *P. tenera* 66.6 and *P. seriata* 67.4). The prevalence of GC-rich trinucleotide seemed to correlate with high GC composition in genus *Pyropia*. However, the correlation between GC content and SSRs type preference cannot be taken as a rule, and species *Mesostigma*, *Tortula* and *Allium* were the typical examples (Stackelberg *et al.*, 2006; Victoria *et al.*, 2011). Thus, the correlation between GC content and SSR type in genus *Pyropia* should be treated with caution.

The origin and functional role of SSRs in expression sequences were still not well understood. Researches have found that the SSR-containing unigenes have a range of functions such as metabolic enzymes, structural and storage proteins, disease signaling, and transcription factors, *etc.*, (Kantety *et al.*, 2002). The SSR-containing unigenes from *Brassica* species were mainly involved in nucleotide or protein binding and enzyme activity, and preferentially functioned in membranes and cytoplasm (An *et al.*, 2011). The SSR-containing unigenes detected in *Camellia sinensis* were homologous to proteins with distinct molecular functions such as binding, catalytic, transport, enzyme regulation, and structural activities in

different biological processes, and cellular and sub-cellular organelles (Sharma *et al.*, 2009). In this research, GO annotation result exhibited that the major SSR-containing unigenes in *P. yezoensis* have the functions of 'metabolic process', 'binding', 'catalytic activity' and existed as the forms of 'cell', 'cell component', and 'organelle'. KEGG result showed that 'signal transduction' was the most important pathway SSR-containing unigenes participated. More work has to be done to explore the relationship between SSRs and the unigenes functions.

SSRs located in coding regions could code amino acids in translation process. The most common amino acid coded by SSRs was Lys in *A. thaliana* and Arg in sugarcane (Li *et al.*, 2004). The dominant amino acids coded by SSRs in *P. yezoensis* were Ala, Arg, Gly and Pro, constituting 82.61% of the total amino acids coded by SSRs loci in ORF region. Ala, Gly and Pro were small amino acids and Arg was hydrophilic amino acid. This pattern was consistent with the conclusions of other researches that small and hydrophilic amino acids were more tolerated than hydrophobic in coding region, while strong selection pressure eliminated codon repeat encoding hydrophobic amino acid in the evolution process (Katti *et al.*, 2001). The component of SSRs would affect the structure and function of the encoded proteins. The instability of amino acids repeats in protein sequences may lead to rapid evolution of new domains in regulatory proteins (Temnykh *et al.*, 2001).

In this research, only 3 (13.04%) out of 23 SSRs loci from transcriptome and 8 (10.39%) out of 77 SSRs loci from genome were validated polymorphic. In the research of *Phaseolus vulgaris*, 31 out of 40 SSRs loci from transcriptome and 26 out of 40 SSRs from genome were polymorphic (Luiz *et al.*, 2007). In kiwi, rice and wheat transcriptomes, 93.5%, 43% and 38% of the SSRs markers were polymorphic respectively (Rajeev *et al.*, 2005). The SSRs in *P. yezoensis* exhibited lower polymorphism characteristic. Many researches exhibited that the polymorphism of SSRs was caused by the strand-slippage in DNA replication process. The SSRs length was the main factor that affects the polymorphism, and longer SSRs sequences would be higher variable (Hans, 2004). From the analysis it was found that the SSRs average sizes in *P. yezoensis* transcriptome and genome were 16.90 bp and 17.23 bp, respectively. Additionally, the size of 94.1% SSRs loci in *Camellia sinensis* unigenes was longer than 20 bp (SSRs selection criterion: length  $\geq 18$  bp for dinucleotide and trinucleotide, SSRs length  $\geq 15$  bp for tetranucleotide, pentanucleotide and hexanucleotide) (Sharma *et al.*, 2009). The SSRs loci average lengths in *C. reinhardtii*, *M. viride*, *Marchantia polymorpha*, *Syntrichia ruralis*, *Physcomitrella patens*, *Selaginella* spp., *Adiantum capillus-veneris*, *Gnetum gnemon*, *Pinus taeda*, *O. sativa* and *A. thaliana* EST databases were 33.21 bp, 34.12 bp, 22.56 bp, 23.84 bp, 24.20 bp, 23.71 bp, 31.14 bp, 23.62 bp, 30.89 bp, 23.44 bp and 26.52 bp respectively (SSRs loci selection criterion: length  $\geq 20$  bp for dinucleotide to hexanucleotide) (Victoria *et al.*, 2011). Using the same selection criterion (SSRs selection criterion: length  $\geq 20$  bp for dinucleotide to hex-

anucleotide), the SSRs average sizes were 22.81 bp in transcriptome and 24.05 bp in genome of *P. yezoensis*. The SSRs length in *P. yezoensis* seemed to be size-restricted. This phenomenon was also found in other plants species. SSRs analysis in *Sorghum bicolor* unigenes showed the dominant SSRs lengths were shorter than 20 bp and only 19.8% markers were found to be polymorphic (Reddy *et al.*, 2012). High GC content feature (67.78 in transcriptome and 63.60 in genome) may be another reason resulting in low polymorphism in *P. yezoensis*. GC-rich SSRs types had lower possibility in strand-slip-page during sequence replication process. Research in rice showed that the GC-rich trinucleotide which contain  $\geq 2G$  and/or C in their repeating units had fewer alleles and lower polymorphism (Temnykh *et al.*, 2001). SSRs from transcriptome were preferable over genomic SSRs as they were associated with unigenes functions and could be used as trait-markers in MAS (An *et al.*, 2011). More SSRs markers obtained from *P. yezoensis* transcriptome should be validated and related to unigenes functions. At the same time, the SSRs markers validated in this research were effective detectors to identify the purity and heterozygous conchocelis.

## 5 Conclusions

This research firstly systemically analyzed the distribution, putative function and polymorphism characteristics of SSRs in *P. yezoensis* transcriptome. Research showed that SSRs were abundant in *P. yezoensis* transcriptome and the trinucleotide was the dominant type. The SSRs showed a GC bases preference. The majority of annotated SSR-containing unigenes have the functions of 'metabolic process', 'binding', 'catalytic activity' and existed as the forms of 'cell', 'cell component', 'organelle'. The dominant amino acids coded by SSRs were Ala, Arg, Gly and Pro. The lower polymorphism of SSRs may be caused by shorter sequences and higher GC content.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 31372517, 31672641), the Scientific and Technological Innovation Project Financially Supported by Qingdao National Laboratory for Marine Science and Technology (No. 2015ASKJ02), the Project of National Infrastructure of Fishery Germplasm Resources (No. 2016DKA30470), Fundamental Research Funds for the Central Universities (Nos. 201762016, 201562018, 201564009), and the Program for Chinese Outstanding Talents in Agriculture Scientific Research.

## References

- An, Z., Gao, C. H., Li, J. N., Fu, D. H., Tang, Z. L., and Ortegón, O., 2011. Large-scale development of functional markers in *Brassica* species. *Genome*, **54**: 763-770.
- Asadi, A. A., and Monfared, S. R., 2014. Characterization of EST-SSR markers in durum wheat EST library and functional analysis of SSR-containing EST fragments. *Molecular Genetic and Genomics*, **289**: 625-640.
- Bi, Y. H., Wu, Y. Y., and Zhou, Z. G., 2014. Genetic diversity of wild population of *Pyropia haitanensis* based on SSR analysis. *Biochemical Systematics and Ecology*, **54**: 307-312.
- Biswas, M. K., Xu, Q., Mayer, C., and Deng, X. X., 2014. Genome wide characterization of short tandem repeat markers in sweet orange (*Citrus sinensis*). *PLoS One*, **9** (8): e104182.
- Choi, S., Hwang, M. S., Im, S., Kim, N., Jeong, W. J., Park, E. J., Gong, Y. G., and Choi, D. W., 2013. Transcriptome sequencing and comparative analysis of the gametophyte thalli of *Pyropia tenera* under normal and high temperature conditions. *Journal of Applied Phycology*, **25**: 1237-1246.
- Hans, E., 2004. Microsatellites: Simple sequences with complex evolution. *Nature Review Genetics*, **5**: 435-445.
- Im, S., Choi, S., Hwang, M. S., Park, E. J., Jeong, W. J., and Choi, D. W., 2015. Denovo assembly of transcriptome from the gametophyte of the marine red algae *Pyropia seriata* and identification of abiotic stress response genes. *Journal of Applied Phycology*, **27**: 1343-1353.
- Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., and Dhawan, A. K., 2011. Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, **177**: 309-334.
- Kantety, R. V., Rota, M. L., Matthews, D. E., and Sorrells, M. E., 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology*, **48**: 501-510.
- Kashi, Y., and King, D. G., 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, **22** (5): 253-259.
- Katti, M. V., Ranjekar, P. K., and Gupta, S. V., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution*, **18** (7): 1161-1167.
- Li, Y. C., Korol, A. B., Fahima, T., and Nevo, E., 2004. Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, **21** (6): 997-1007.
- Liu, F. L., Hu, Z. M., Liu, W. H., Li, J. J., Wang, W. J., Liang, Z. R., Wang, F. J., and Sun, X. T., 2015. Distribution, function and evolution characterization of microsatellite in *Sargassum thunbergii* (Fucales, Phaeophyta) transcriptome and their application in marker development. *Scientific Reports*, **6**: 18947.
- Luiz, R. H., Tatiana, C., Luis, E. A. C., Luciana, L. B., Anete, P. S., Maeli, M., Sergio, A. M. C., Alisson, F. C., Luciano, C., Eduardo, F. F., Marcos, V. B. M. S., Siu, M. T., and Maria, L. C. V., 2007. Development, characterization, and comparative analysis of polymorphism at common bean SSR loci isolated from genic and genomic sources. *Genome*, **50**: 266-277.
- Mun, J. H., Kim, D. J., Choi, H. K., Gish, J., Debelle, F., Mudge, J., Denny, R., Endré, G., Saurat, O., Dubez, A. M., Kiss, G. B., Roe, B., Young, N. D., and Cook, D. R., 2006. Distribution of microsatellites in the genome of *Medicago truncatula*: A resource of genetic markers that integrate genetic and physical maps. *Genetics*, **172**: 2541-2555.
- Rajeev, K. V., Andreas, G., and Mark, E. S., 2005. Genic microsatellite markers in plants: Features and applications. *Trends in Biotechnology*, **23** (1): 48-55.
- Reddy, R. N., Madhusudhana, R., Mohan, S. M., Chakravarthi, D. V. N., and Seetharama, N., 2012. Characterization development and mapping of unigene-derived microsatellite markers in sorghum. *Molecular Breeding*, **29**: 543-564.
- Sharma, R. K., Bhardwaj, P., Negi, R., Mohapatra, T., and Ahuja, P. S., 2009. Identification, characterization and utilization of

- unigene derived microsatellite markers in tea (*Camellia sinensis* L.). *BMC Plant Biology*, **9**: 53.
- Stackelberg, M., Rensing, S. A., and Reski, R., 2006. Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biology*, **6**: 9.
- Sun, P. P., Mao, Y. X., Li, G. Y., Cao, M., Kong, F. N., Wang, L., and Bi, G. Q., 2015. Comparative transcriptome profiling of *Pyropia yezoensis* (Ueda) M. S. Hwang & H. G. Choi in response to temperature stresses. *BMC Genomics*, **16**: 463.
- Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, **11**: 1441-1452.
- Tóth, G., Gáspári, Z., and Jurka, J., 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research*, **10**: 967-981.
- Victoria, F. C., Maia, L. C., and Oliveira, A. C., 2011. *In silico* comparative analysis of SSR markers in plants. *BMC Plant Biology*, **11**: 15.
- Wang, H. B., Jiang, J. F., Chen, S. M., Qi, X. Y., Peng, H., Li, P. R., Song, A. P., Guan, Z. Y., Fang, W. M., Liao, Y., and Chen, F. D., 2013. Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale unigene assembly and SSR marker discovery. *PLoS One*, **8** (4): e62293.
- Xie, C. T., Li, B., Xu, Y., Ji, D. H., and Chen, C. S., 2013. Characterization of the global transcriptome for *Pyropia haitanensis* (Bangiales, Rhodophyta) and development of cSSR markers. *BMC Genomics*, **14**: 107.
- Yan, X. H., and Aruga, Y., 2000. Genetic analysis of artificial pigmentation mutants in *Porphyra yezoensis* Ueda (Bangiales, Rhodophyta). *Phycological Research*, **48**: 177-187.
- Yoji, N., Naobumi, S., Masahiro, K., Nobuhiko, O., Motoshige, Y., Yuya, S., Masataka, S., Yoshiya, F., Koji, S., Atsumi, T., Takanori, K., Ichiro, N., Fuminari, I., Kazuhiro, N., Motohiko, S., Tokio, W., Satoru, K., Kiyoshi, I., Takashi, G., and Kazuho, I., 2013. The first symbiont-free genome sequence of marinered alga, susabi-nori (*Pyropia yezoensis*). *PLoS One*, **8** (3): e57122.

(Edited by Qiu Yantao)