Mining of EST-SSR markers in clam *Meretrix meretrix* larvae from 454 shotgun transcriptome

Hongxia Wang, Pin Huan, Xia Lu and Baozhong Liu*

Key Laboratory of Experimental Marine Biology, Institute of Oceanology,
Chinese Academy of Sciences, Qingdao 266071, China

(Received 24 February 2011, accepted 4 June 2011)

A total of 2,970 EST-SSRs (2.38%) were identified by transcriptome sequencing of clam *Meretrix meretrix* (751,970 reads, ~310.82 Mbp), using 454 Genome Sequencer FLX next-generation sequencing platform. Dinucleotide SSR was the dominant repeat type (40.2%), followed by trinucleotide (37.8%), tetranucleotide (12.0%) and pentanucleotide (2.0%) SSR. The dominant repeat motif was AT (71.3%) in the dinucleotide SSR type and AAC (45.6%) in the trinucleotide SSR type. Nearly 79% of all microsatellites had flanking sequences suitable for PCR primer design. Half of PAL were found to be polymorphic in a subset of 40 primer pairs randomly selected. Specifically, the density of dinucleotide, trinucleotide and tetranucleotide repeats showed significant variation among four development stages (trochophore, D-veliger, pediveliger and postlarva). The results suggested that dinucleotide, trinucleotide and tetranucleotide SSRs may play an important role in contributing to the different expression profiles in larval stages.

Key words: high-throughput marker identification, EST-SSR, developmental stages, *Meretrix meretrix*

INTRODUCTION

Microsatellites or simple sequence repeats (SSRs) are highly polymorphic sequences distributed throughout the genome. SSR markers are co-dominant, multi-allelic and easy for scoring, with a wide range of applications including genetic mapping, quantitative trait loci (QTL) association, kinship analysis, population genetics, and evolutionary studies (Pérez et al., 2005). Although their applications in genetic analysis have already been widely demonstrated, the traditional approaches to develop microsatellite loci from non-model species require considerable investment. SSR isolation relies on the creation and screening of enriched microsatellite libraries including cloning and hybridization to detect positive clones, plasmid isolation and Sanger sequencing.

The introduction of a massively parallel pyrosequencing technology developed by 454 Life Sciences Technology has made sequencing cheaper and more efficient. This new technology has been widely applied to the sequencing of microbial genomes, genotyping, genome resequencing, transcriptome profiling and methylation studies (Sithichoke et al., 2009). Fast identification of microsatellites from genome shotgun sequences has been recently acknowledged and investigated in several species, e.g.

copperhead snake (*Agkistrodon contortrix*) (Castoe et al., 2010), blue duck (*Hymenolaimus malacorhynchus*) (Abdelkrim et al., 2009) and mungbean (*Vigna radiata*) (Sithichoke et al., 2009).

Most of the markers developed by genome shotgun sequences correspond to type II markers lacking known functions (Weber, 1990). Similar to genome DNA, EST sequences also contain SSR sequences, which can be used to develop SSR markers (Liu et al., 1999; Whan et al., 2000; Eujayl et al., 2002; Karsi et al., 2002). Therefore, mining microsatellite sequences from the EST sequences of the transcriptome is readily accessible and can be immediately used in the development of specific markers such as EST-SSRs, consuming less time and cost. Furthermore, EST-SSRs, as rich resources of type I markers associated with genes of known functions, can be more useful for comparative gene mapping (Liu et al., 1999) and physical mapping.

The clam *M. meretrix* is an important commercial bivalve and can be widely found in the coastal and estuarine areas of South and Southeast Asia (Tang et al., 2006). Like other bivalves, the production of *M. meretrix* mainly relies on the collection of natural spat, despite their economical importance. Recently, selective breeding programs have been initiated in clam *M. meretrix*. The quantitative genetic studies of economical characters such as growth suggested that significant improvement could be achieved by selective breeding (Wang et al.,

Edited by Ali Masoudi-Nejad

* Corresponding author. E-mail: bzliu@qdio.ac.cn

2011). Therefore, the development of genetic and genomic data are likely to contribute to the development of selective breeding programs, and, more generally, to further the understanding about the genome of this species. In light of all these, we generated and characterized a total of 751,970 original reads covering 310.82 Mb of the *M. meretrix* larvae transcriptome profiling using 454 sequencing technology. The objective of this study was to identify EST-SSR markers and investigate the type and distribution of repeat motifs in different larval stages of *M. meretrix*. The results will facilitate the use of molecular markers in *M. meretrix* genetic analysis and breeding. Moreover, the EST-SSRs specific to different larval stages were compared, which would be of great value to the understanding of the repeat function of gene expression and to the studies of larval development.

MATERIALS AND METHODS

Preparation of cDNA libraries for 454 sequencing

The *M. meretrix* larvae in four developmental stages, i.e., trochophore, D-veliger, pediveliger and postlarva, were used as the source of cDNA libraries in this study. The larvae were collected from a hatchery in July 2009. Then they were frozen in liquid nitrogen then stored at -80°C till use. Samples of *M. meretrix* larvae in each larval stage were used for RNA purification and sequencing. Total RNA was extracted using phenol-chloroform extraction, and mRNA was purified through affinity chromatography. Four microgram cDNA were synthesized by SMART cDNA synthesis method. A total of four cDNA libraries were ultimately sequenced using 454 Life Sciences technology on the Genome Sequencer (GS) FLX System.

454 Shotgun sequencing of *M. meretrix* transcriptome

A total of 751,970 reads were generated with the average read length of 413.35 bp covering 310.82 M. They were obtained from the samples of *M. meretrix* larvae in four larval stages (trochophore, D-veliger, pediveliger and postlarva). All the original reads have been deposited in the NCBI Sequence Read Archive (SRA, www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) (Accession No.: SRA021052). A total of 35,205 contigs were yielded with 89,532 reads remaining at singletons after the assembly. Further analysis revealed that 94.66% of the unique sequences showed a comparatively lower expression level (depth < 10), indicating a deep-coverage sequencing (Huan et al., 2011).

Mining of EST-SSR loci In order to identify microsatellite markers, non-redundant sequences were screened for SSRs using MISA software (<http://pgrc.ipk-gatersleben.de/misa>). In the search for SSR standard, we defined SSRs as dinucleotide repeat (DNP) ≥ 12 bases; tri-

nucleotide repeat (TNP) ≥ 15 bases; tetranucleotide repeat (TTNP) ≥ 20 bases; pentanucleotide repeat (PNP) ≥ 20 base; hexanucleotide repeat (HNP) (and more) ≥ 24 bases (Cardle et al., 2000). Reverse-complement repeat motifs and translated or shifted motifs were grouped together (e.g. AC representing AC, CA, TG and GT) due to the double-stranded nature of DNA and the fact that the start site of a SSR could be considered arbitrary (Jurka and Pethiyagoda, 1995). Primer pairs were designed to amplify microsatellite regions using BatchPrimer3 (Frank et al., 2008). The compound repeat type microsatellite sequences were not included in the designing of amplification primers. We termed such loci with PCR primer as potentially amplifiable loci or PAL.

Statistical analysis The expression level of EST-SSRs was considered in the evaluation of the SSR distribution in larval stages. The total number of EST-SSRs in each developmental stage (trochophore, D-veliger, pediveliger and postlarva) was calculated using the following formula:

$$N_{\text{SSR}} = \sum_{i=1}^n 1 \times D(i), \text{ where } D(i) \text{ is the depth of each contig}$$

and singleton containing the SSRs. The depth of contigs containing SSRs ranged between 2 and 375, and the depth of singletons was one. The same formula $N_{\text{T}} = \sum_{i=1}^n 1 \times D(i)$ was also used in the evaluation of the total number of unique sequences, where $D(i)$ is the depth of all contigs and singletons. The depth of contigs ranged between 2 and 1613.

Chi-square (χ^2) goodness-of-fit tests with three degrees of freedom were applied to test whether SSR density was significantly different in four larval stages. The expected number of SSRs was calculated using the following formula: $E_i = N_{\text{SSR}}/L \times L_i$ (Lawson and Zhang, 2008), based on the assumption that the SSR density was the same in the four larval stages. In this formula, N is the total number of SSRs in the four larval stages; E_i is the expected number of SSRs in the four larval stages, trochophore, D-veliger, pediveliger and postlarva, respectively; L is the total length of base pairs in the four larval stages; and L_i is the length of base pairs in the four larval stages, trochophore, D-veliger, pediveliger and postlarva, respectively.

RESULTS

The distribution of EST-SSR type in *M. meretrix* transcriptome

2,970 EST-SSRs (2.38%) were isolated from a total of 124,737 contigs and singletons from the above transcriptome, using the Troll software. Among the 2,970 identified EST-SSRs, 1,126 EST-SSRs were identified from contigs and 1,844 EST-SSRs from singletons. These EST-SSRs could be classified into 1194 dinucleotide repeats, 1123 trinucleotide repeats, 357

tetranucleotide repeats, 60 pentanucleotide repeats and 236 other motif types of SSRs including hexanucleotide repeats or repeats with more than 6 bases and compound repeat types (Fig. 1).

Except for 233 compound repeat type loci, 2343 PAL loci had flanking regions and could be used for primer design, proving themselves to be promising candidates for PCR amplification. EST-SSRs containing PAL accounted for 78.9% of all microsatellites. The trinucleotide repeats had the most PAL (979), accounting for 33% of the total EST-SSRs. Within each repeat class, besides the compound repeat type, four repeat types (DNP, TNP, TTNP and PNP) had the percentage of PAL higher than 78%

Fig. 1. Frequency distribution of different repeat types of SSR, and subset of these containing PAL (gray), which were identified in Uni-sequences (214,737 contigs and singletons) of clam *M. meretrix*. The numbers on the bars indicate the percentage of each repeat type of SSR in total number.

(Fig. 1).

Through the comparison across the five classes of repeats, dinucleotide SSR was the dominant repeat type (40.2%). TNPs were found in 1123 SSR loci (37.8%), which was close to the number of DNPs, followed by tetranucleotide (12.0%) and pentanucleotide (2.0%) SSR. There were significant differences in the relative abundances of specific repeat motifs (Fig. 2). The most common motif type of DNPs was TA/AT (71.3% of DNPs), followed by AC/TG (18.3% of DNPs) and TC/AG (10.4% of DNPs). The GC/CG motif was not found in the data set. TA/AT motif also had the largest number of PAL loci in the dinucleotide repeat. Trinucleotide repeat motifs were dominated by TTG repeat, which was found in 512 loci (45.6% of TNPs). TTG repeats contained the largest number of PAL loci in TNPs. The least frequent TNP motif were GC-rich (CAG/TGG/TGC), and the accumulative total value was no more than 9% of TNPs. The number of loci identified for TTNPs was 357, which was less than one third of that of DNPs or TNPs. For tetranucleotides, AAAC, ACAG and TGAT were the three most common motifs, half of all the TTNP motifs were observed fewer than 30 times (data not shown).

EST-SSR repeat size in *M. meretrix* transcriptome

The repeat number of motif mainly ranged between 5 and 50. There were 909 SSRs with 6 tandem repeats in DNPs, TNPs and TTNPs, followed by 5 tandem repeats in TNPs (623). The two most common repeat number accounted for 51.6% of total EST-SSRs, followed by 7 tandem repeats (15.1%, 450) and 8 tandem repeats (6.9%, 204). More details about different repeat motifs of di-, tri- and tetranucleotide distribution in EST-SSRs were listed in Fig. 3. It showed that with the increase of repetition number, the number of microsatellites declined

Fig. 2. Frequency distribution of 4 di-nucleotide repeat motifs (A) and 7 tri-nucleotide repeat motifs (B) in Uni-sequences (214,737 contigs and singletons) of clam *M. meretrix*. The subset of these containing PAL is denoted in gray. The numbers on the bars indicate the percentage of each di-nucleotide repeat motif in all di-nucleotide repeat types and each tri-nucleotide repeat motif in all tri-nucleotide repeat types, respectively.

Fig. 3. The distribution of the number of perfect tandem repeat units for di-nucleotide repeats (A), tri-nucleotide repeats(B) and tetra-nucleotide repeats (C). The numbers on the bars indicate the percentage of repeat numbers in each repeat type.

Table 1. EST-SSR densities in different larval stages of *M. meretrix*

	Bases	N _T	N _{SSR}			Di-N _{SSR}			Tri-N _{SSR}			Tetra-N _{SSR}			Penta- N _{SSR}			Hexa- N _{SSR}	
			Obs	Exp	χ ²	Obs	Exp	χ ²	Obs	Exp	χ ²	Obs	Exp	χ ²	Obs	Exp	χ ²	Obs	Exp
Trochophore	14486080	155623	5500 (3.5%)	4112		1524	1340		3279	2229		632	500		34	33		31	9
D-veliger	15379670	157091	4183 (2.7%)	4365		1366	1423		2049	2367		743	531		25	35		0	9
Pediveliger	14144093	123827	3321 (2.7%)	4015		1255	1309		1632	2176		393	488		40	33		1	8
Postlarva	12944013	116551	3162 (2.7%)	3674		1125	1198		1804	1992		199	447		32	30		2	8
	56953856	553092	16166		667.42**	5270		34.22**	8764		691.08**	1967		275.43**	131		4.51	34	

N_T: the total number of unique sequences.
N_{SSR}: the total number of unique sequences containing SSR.

rapidly. Variations in EST-SSRs in most cases were due to the variations in repeat number. When slippage mutations happen, expansion occurs more frequently to short microsatellites and contraction occurs more frequently to long microsatellites. And the length-dependent mutation pattern explained the scarcity of long microsatellites (Weber, 1990; Temnykh et al., 2000).

EST-SSR distribution in transcriptome in different larval stages The transcriptome in each larval stage was also generated, respectively. The proportion of EST-SSRs in D-veliger, pediveliger, postlarva stage was similar, being 2.7%, and that in the trochophore stage was 3.5% (Table 1). However, the Chi-square goodness-of-fit

test showed that there were considerable variations in the density of EST-SSRs in the four developmental stages (Table 1). The densities of di-nucleotide repeats in D-veliger, pediveliger and postlarva stages were slight lower by 4%–6% than expected density, whereas in trochophore stage the density of di-nucleotide repeats was higher by 13.7% than expected density. The density of tri-nucleotide repeats showed the same tendency as di-nucleotide repeats in the four developmental stage, but with more significant variation, e.g., tri-nucleotide repeats density in the trochophore stage was higher by 47.1% than expected density. For the tetra-nucleotide, the repeats densities in the trochophore and D-veliger stage were higher by 26.4% and 39.9% than expected density, respec-

tively, while in pediveliger and postlarva stage, the repeat density was significantly lower than expected density, especially in postlarva stage, it was lower by 55.5% than expected density. The density of EST-SSRs showed no significant variation in four developmental stages for pentanucleotide repeat. The total number of hexanucleotide repeats was small, and therefore it was not analyzed in the goodness-of-fit test.

Development and evaluation of EST-SSR markers From each SSR type, we randomly selected a subset of 40 primer pairs including 6 dinucleotides, 11 trinucleotides, 19 tetranucleotides and 4 pentanucleotides to assess the PAL availability rate. Approximately 65% of PAL gave PCR products of expected size and half of PAL were found to be polymorphic (Table 2). The trinucleotides SSR showed the highest PAL availability rate (72.72%), followed by dinucleotides (66.67%) and tetranu-

Table 2. Primer sequences, characteristics (type and number of repeats), amplification conditions (optimized annealing temperature) and Polymorphic (Yes or No) in the 40 microsatellite loci developed from ESTs in *M. meretrix*

Locus	Primer sequences (5'-3') (F, forward; R, reverse)	Motif	Type	Ta (°C)	Size (bp)	TSA Accession	Polymorphic
<i>MM2034</i> *	F: CCACCAAGTCTACAGTCTTCA R: CAAAAAGGACCAAAACAAAAGT	(GA) _n	di	55	153	JI259636	Y
<i>MM6277</i> *	F: GACCAATGAAATTCAGTCAAA R: TAGAAAACATAAGGCACTGCT	(TG) _n	di	55	155	JI263848	Y
<i>MM7501</i>	F: TTTCTGTGCTTGATTGAAGTT R: AACCACCTTTTACACAGAATGC	(TG) _n	di	55	none	JI265066	N
<i>MM8777</i>	F: ACTCTGCCTTGTGCTTTTTAT R: CTTATTCAGGGAGGACTTGAT	(AT) _n	di	55	167	JI266337	N
<i>MM9504</i>	F: TATGACTGACACACACGCATA R: GACGATTATTTCACGTAAAGG	(TG) _n	di	55	152	JI267057	Y
<i>MM16268</i> *	F: TCGGATGATTGCCCTTAC R: ATTCCCTGTGCTTTACCC	(AC) _n	di	49	150	JI273780	Y
<i>MM98</i> *	F: AGCGAAGATTTTAACAAA R: ATCTTCAACTCACCATCATCA	(TGA) _n	tri	55	157	JI257709	Y
<i>MM573</i>	F: TTAACAACTTGCATTTCTC R: ATCATCATCTTCAACTACCA	(TGA) _n	tri	55	146	JI258183	Y
<i>MM1031</i>	F: GGTGTGAAAGACAATTGAGA R: TGCCCATTAAGACAAGACTA	(CAT) _n	tri	55	152	JI258639	Y
<i>MM1272</i> *	F: AGGCCTGAAATGTTTCTTAAC R: AAACGACGCTCAACATCTAT	(TTG) _n	tri	55	150	JI258879	Y
<i>MM3246</i>	F: GTTACTGAGCCTGTGAAGATG R: GATGTCTGTGCCTACCACTT	(TTG) _n	tri	55	none	JI260839	N
<i>MM5138</i>	F: TTAACAACTTGCATTTCTC R: TCATCATCTTCAACTCACCAT	(TGA) _n	tri	55	150	JI262714	N
<i>MM6477</i> *	F: CTGATTGAAATCTTCGGTATG R: GACTAGGGGATTAGGTCTTGA	(CAT) _n	tri	55	148	JI264048	Y
<i>MM8761</i>	F: CTCATCATATTCATGCTCCTC R: TGAAGAAAAATGAAGTCGGTA	(TCA) _n	tri	55	none	JI266321	N
<i>MM12736</i> *	F: GTCAGCGAAGATTTTAACAAA R: TCATCATCTTCAACTCACCAT	(TGA) _n	tri	55	150	JI270267	Y
<i>MM26715</i> *	F: ACATCATCATCTCAACTCACC R: GTCAGCGAAGATTTTAACAAA	(ATC) _n	tri	57	154	JI284180	Y
<i>MM12295</i> *	F: AAATGGTTCCTAACGTTTCATT R: AAGCTTGTATAAGTGACGAG	(ACA) _n	tri	55	161	JI269829	Y

Continued

Table 2. Continued

<i>MM1642</i>	F: GAACATTGTTTCTGTTTGTCC R: TGTACATCGGCATTAAGTTTT	(TGTC) _n	tetra	55	none	JI259246	N
<i>MM2016*</i>	F: AAGCCTCATGTATTGAACAAA R: ATGCACAATTTCTGAACAAC	(TGTC) _n	tetra	55	148	JI259618	Y
<i>MM2525</i>	F: GCACTTTCTATCTTTGGCTTT R: TCAGTTACAAAAAGTGCTGTG	(TCCA) _n	tetra	55	none	JI260124	N
<i>MM3050</i>	F: TATGCCCAAGATAACCATAGA R: GAATGAATGAATGGATGAATG	(TTCA) _n	tetra	55	none	JI260645	N
<i>MM3923*</i>	F: TTTTCGTCTTAATGAGGGTTA R: GTTTGTGAAATAGTGCTCTGC	(AATC) _n	tetra	55	154	JI261510	Y
<i>MM5358*</i>	F: TTCTACTGACCTAAGCTGCTG R: CCATATGTGTCATTGGAAGTT	(AATC) _n	tetra	55	149	JI262933	Y
<i>MM5394</i>	F: CTTTCAATTTGGACAGAACC R: CTAGCGGAGAAATTGGTAAGT	(TTTG) _n	tetra	55	none	JI262969	N
<i>MM6717*</i>	F: TGGAAGTTGATAATCTGAAGC R: TTCTACTGACCTAAGCTGCTG	(TTGA) _n	tetra	55	151	JI264287	Y
<i>MM6863*</i>	F: AACATCCAAGTACAGTTGCTG R: TTTACAGACAGACAGGCAGAT	(TGTC) _n	tetra	55	152	JI264433	Y
<i>MM7226</i>	F: GCACTTTCTATCTTTGGCTTT R: TCAGTTACAAAAAGTGCTGTG	(TCCA) _n	tetra	55	none	JI264793	N
<i>MM31246</i>	F: TGAATGATGAAGTGAAGTGAA R: TATAAATCAATTGCACCGAAG	(CTCA) _n	tetra	55	none	JI288684	N
<i>MM35039</i>	F: CTGGTCTTCCCAAATAGTTCT R: ATTTGAAGCGTGTGTTGTATT	(TGAA) _n	tetra	55	243	JI292449	N
<i>MM26863*</i>	F: AAGCCTCATGTATTGAACAAA R: ATGCACAATTTCTGAACAAC	(TGTC) _n	tetra	50	127	JI284326	Y
<i>MM7583</i>	F: CCATGAATTCAGACTTGTTTT R: TTAAGGTCAAGATCGCAGTAG	(AATC) _n	tetra	55	none	JI265148	N
<i>MM7721</i>	F: AAGAGAGTGAAGGAAGCCTTA R: ATGCACAATTTCTGAACAAC	(TGTC) _n	tetra	55	145	JI265286	N
<i>MM8105*</i>	F: AGTTGCCTTGAAGTAAAGTCC R: CATGATCAATCATTGGTTACA	(TGAT) _n	tetra	55	151	JI265669	Y
<i>MM8365*</i>	F: AAGCAAACTATCACATTCCA R: AGTCTGTCATCTCCTTTTCGAT	(ACAG) _n	tetra	55	150	JI265926	Y
<i>MM8526</i>	F: CTGTCTGTTTGCTTGTCTCTC R: CAATCGCTTCAGTGAAAAGTTA	(TCTG) _n	tetra	55	none	JI266086	N
<i>MM13589</i>	F: CGCAGTAACCTAATTGACTTG R: GTGACAATCACTGAATCCATC	(TTGA) _n	tetra	55	none	JI271112	N
<i>MM5584</i>	F: CAGCTGTTATTTTCATAATGTGAC R: GCGGTACATGTTATTGTTTGT	(CTGAA) _n	penta	55	none	JI263159	N
<i>MM7561</i>	F: AGAAGACTTAGTTGGCACCAC R: CCAACAGTGAAGATATTGCAT	(ATCTG) _n	penta	55	none	JI265126	N
<i>MM14624</i>	F: CATGCAACTACATCACCTTTT R: CCCACGACCATTATCTTCTA	(TTCTA) _n	penta	55	243	JI272144	N
<i>MM21942</i>	F: ATAATGCCCAAACAAGTTCAT R: ATTCGACATTTTGATTGACAG	(TACAG) _n	penta	55	152	JI279433	N

Note: the loci with * quote from Lu et al. 2011.

cleotides (42.11%) types (Table 2). No polymorphic locus was detected in the 4 pentanucleotides (Table 2). Beside three loci (*MM9504*, *MM573* and *MM1031*) significantly deviated from HWE ($P < 0.001$) were excluded from the genetic analysis, the characterization of 17 polymorphic EST-SSRs was evaluated in 30 individuals randomly selected from a natural population in Northern China (Lu et al., 2011).

DISCUSSION

Using the next-generation sequencing technology, we obtained more than 700,000 high-quality reads, which could be assembled into 124,737 unique sequences. Although some genomic SSRs of *M. meretrix* have been developed based on standard enrichment-based approaches, its small quantity still limits its application in genetic mapping and MAS (marker assisted selection). Here an extensive set of 2,970 microsatellites over 12 bp was identified from *M. meretrix* EST dataset without previously knowing anything about the target genome sequence. The rate of successful amplification of EST-SSRs was high in this batch of PAL, and their primers were available for evolutionary and population genetic research as well as high-resolution chromosome linkage mapping studies of *M. meretrix*.

The frequency data of SSRs in protein-encoding or non-coding DNA have been widely reported in vertebrates (Moran, 1993; Jurka and Pethiyagoda, 1995; Edwards et al., 1998; Castoe et al., 2010) and plants (Wang et al., 1994; Varshney et al., 2002; Zeng et al., 2010). For example, EST-SSR density of dicotyledonous species ranged between 2.65% and 16.82% (Kumapala and Mukhopadhyay, 2005), very few studies were conducted on the bivalve. Cruz et al. (2005) once concatenated data from 3165 DNA sequences of 326 bivalve species, accounting for 101 Mb, to determine taxonomic patterns of microsatellite distribution in genomic regions of markedly different functionality. They found that the microsatellite density was 2.9 times higher in introns than in exons. Tanguy et al. (2008) found that less than 0.5% of EST-SSRs were identified in four bivalve species, *Bathymodiolus azoricus*, *Crassostrea gigas*, *Mytilus edulis* and *Ruditapes decussatus*. The frequency of EST-SSRs detected in *M. meretrix* was much higher than that of the bivalve species studied. However, it was in agreement with the findings by Li et al. (2010), who identified 3.1% of microsatellite-containing sequences in 1796 unigenes of *M. meretrix*. The species difference might be the main reason for the variations across species. The EST-SSR frequency is also dependent on other factors, such as different software used in detecting SSRs and customized parameters including the repeat length threshold and the number of repeat unit in microsatellites. For example, some SSR identifying tools can detect imperfect and compound

SSRs, while others can only identify perfect SSRs or perfect and compound SSRs (Sharma et al., 2007). In the present study, mononucleotide repeat motifs were excluded in our strategy for identifying microsatellites. The most dominant motifs were di- and tri-nucleotides, which totally accounted for 78.0% of EST-SSRs. It was not consistent with the results reported for other clam species (Tanguy et al., 2008). This result also suggested that there was significant heterogeneity in the most dominant type of repeats, i.e. di- vs tri- nucleotide across bivalve species (Tanguy et al., 2008). In addition, it showed that the most abundant motif in *M. meretrix* was TA/AT, and the ratio of the (TA/AT)_n motif to the second most abundant motif (in our case (AC/TG)_n) was nearly 4:1. These data were completely contrary to those of *M. galloprovincialis* in the distribution type of dinucleotide motif (Cruz et al., 2005). The great variance in motif abundance among species has been reported in bivalves (Cruz et al., 2005; Tanguy et al., 2008).

In all vertebrates, GC-rich motif of trinucleotide is common in exons, but they are less common in intronic sequences (Toth et al., 2000). It was interesting that there was no CG/GC repeat motif and very few CCG/CGG repeats in our results. The rarity of the CCG/CGG repeat units had been reported in a large number of plants (Kumapala and Mukhopadhyay, 2005). Selection might also be against the CCG/CGG repeat unit due to the requirements of splicing mechanism (Li et al., 2004). Long CCG/CGG sequences was competitive against splicing machinery components, thus giving rise to inadequate splicing. Moreover, CCG/CGG repeats might form potential higher structures such as hairpin and quadruplex, and thus affects the efficiency and accuracy of splicing and the formation of mature mRNA (Coleman and Roesser, 1998; Toth et al., 2000).

The larvae of *M. meretrix* undergoes a planktonic stage including trochophore, D-veliger, pediveliger, and then metamorphose into postlarval stage. Furthermore, trochophores live on yolks and the larvae start to feed since D-veliger stage. A preliminary differential analysis revealed that several unique sequences showed very different expression patterns in different stages (Huan et al., 2011). It is unclear why di-, tri- and tetra- nucleotides EST-SSR densities showed significant variation in four developmental stages. Microsatellites were non-randomly distributed throughout eukaryotic genomes, showing different properties in genomic regions of different functionality (Katti et al., 2001). Moreover, more and more studies suggested that the microsatellites, which were regarded as 'junk DNA' in the past, may play a more active role in affecting gene expression (Toutenhoofd et al., 1998; Kenneson et al., 2001; Persico et al., 2006; Lawson and Zhang, 2008). The knowledge of the patterns of microsatellite distribution may help us understand its role in gene regulation and its evolutionary properties. More

researches on microsatellite distribution combined with the gene expression profiles in larval stages of *M. meretrix* need to be done to validate its biological functions.

In conclusion, the molecular markers, such as EST-SSRs, can be applied in the research fields of biodiversity, taxonomy, and population genetics etc. This EST dataset of *M. meretrix* larvae specific to different developmental stages is a powerful resource for developmental biology and genetics studies.

This project was financially supported by National Basic Research Program of China (2010CB126403) and the Key Program of National Natural Science Foundation of China (30730071).

REFERENCES

- Abdelkrim, J., Robertson, B. C., Stanton, Jo-Ann., L., and Gemmell, N. J. (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Bio Techniques* **46**, 185–192.
- Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* **156**, 847–854.
- Castoe, T. A., Polle, A. W., Gu, W., de Koning, A. P. J., Daza, J. M., Smith, E. N., and Pollock, D. D. (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources* **10**, 341–347.
- Coleman, T. P., and Roesser, J. R. (1998) RNA secondary structure: an important *cis*-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry* **37**, 15941–15950.
- Cruz, F., Pérez, M., and Presa, P. (2005) Distribution and abundance of microsatellites in the genome of bivalves. *Gene* **346**, 241–247.
- Edwards, Y. J., Elgar, G., Clark, M. S., and Bishop, M. J. (1998) The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analysis. *Journal of Molecular Biology* **278**, 843–854.
- Eujayl, I., Sorrells, M. E., Baum, M., Wolters, P., and Powell, W. (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics* **104**, 399–407.
- Frank, M. Y., Huo, N., Gu, Y. Q., Luo, M., Ma, Y., Hane, D., Lazo, G. R., Dvorak, J., and Anderson, O. D. (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**, 253, doi:10.1186/1471-2105-9-253.
- Huan, P., Wang, H., and Liu, B. (2011) Transcriptomic analysis of the clam *Meretrix meretrix* on different larval stages. *Marine Biotechnology* doi: 10.1007/s10126-011-9389-0.
- Jurka, J., and Pethiyagoda, C. (1995) Simple repetitive DNA sequences from primates: compilation and analysis. *Journal of Molecular Evolution* **40**, 120–126.
- Karsi, A., Cao, D., Li, P., Patterson, A., Kocabas, A., Feng, J., Ju, Z., Mickett, K. D., and Liu, Z. (2002) Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. *Gene* **285**, 157–168.
- Katti, M. V., Ranjekar, P. K., and Gupta, V. S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution* **18**, 1161–1167.
- Kenneson, A., Zhang, F. P., Hagedorn, C. H., and Warren, S. T. (2001) Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Human Molecular Genetics* **10**, 1449–1454.
- Kumpatla, S. P., and Mukhopadhyay, S. (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* **48**, 985–998.
- Lawson, M. J., and Zhang, L. (2008) Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. *Gene* **407**, 54–62.
- Li, H., Liu, W., Gao, X., Zhu, D., Wang, J., Li, Y., and He, C. (2010) Identification of host-defense genes and development of microsatellite markers from ESTs of hard clam *Meretrix meretrix*. *Molecular Biology Reports* doi: 10.1007/s11033-010-0165-4.
- Li, Y. C., Korol, A. B., Fahima, T., and Nevo, E. (2004) Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution* **21**, 991–1007.
- Liu, Z., Karsi, A., and Dunham, R. A. (1999) Development of polymorphic EST markers suitable for genetic linkage mapping of catfish. *Marine Biotechnology* **1**, 437–447.
- Lu, X., Wang, H., Dai, P., and Liu, B. (2011) Characterization of EST-SSR and genomic-SSR markers in the clam, *Meretrix meretrix*. *Conservation Genetics Resources* doi: 10.1007/s12686-011-9426-3.
- Moran, C. (1993) Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. *Journal of Heredity* **84**, 274–280.
- Pérez, F., Ortiz, J., Zhinaula, M., Gonzabay, C., Calderón, J., and Volckaert, F. A. M. J. (2005) Development of EST-SSR markers by data mining in three species of shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*. *Marine Biotechnology* **7**, 554–569.
- Persico, A. M., Levitt, P., and Pimenta, A. F. (2006) Polymorphic GGC repeat differentially regulates human reelin gene expression levels. *Journal of Neural Transmission* **113**, 1373–1382.
- Sharma, P. C., Grover, A., and Kahl, G. (2007) Mining microsatellites in eukaryotic genomes. *Trends in Biotechnology* **25**, 490–498.
- Sithichoke, T., Prakrit, S., Pichahpuk, U., Juntima, C., Duangjai, S., Worapa, S., Warunee, S., Somvong, T., and Peerasak, S. (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). *BMC Plant Biology* **9**: 137, doi: 10.1186/1471-2229-9-137.
- Tang, B. J., Liu, B. Z., Wang, G. D., Zhang, T., and Xiang, J. H. (2005) Effects of various algal diets and starvation on larval growth and survival of *Meretrix meretrix*. *Aquaculture* **254**, 526–533.
- Tanguy, A., Bierne, N., Saavedra, C., Pina, B., Bachère, E., Kube, M., Bazin, E., Bonhomme, F., Boudry, P., and Boulo, V., et al. (2008) Increasing genomic information in bivalves through new EST collections in four species: Development of new genetic markers for environmental studies and genome evolution. *Gene* **408**, 27–36.
- Temnykh, S., Park, W. D., Ayres, N., Cartinhour, S., Hauck, N., Lipovich, L., Cho, Y. G., Ishii, T., and McCouch, S. R. (2000)

- Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theoretical and Applied Genetics **100**, 697–712.
- Toth, G., Gaspari, Z., and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research **10**, 967–981.
- Toutenhoofd, S. L., Garcia, F., Zacharias, D. A., Wilson, R. A., and Strehler, E. E. (1998) Minimum CAG repeat in the human calmodulin-1 gene 5'untranslated region is required for full expression. Biochimica et Biophysica Acta - Gene Structure and Expression **1398**, 315–320.
- Varshney, R. K., Thiel, T., Stein, N., Langridge, P., and Graner, A. (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular & Molecular Biology Letters **7**, 537–546.
- Wang, H., Chai, X., and Liu, B. (2011) Estimation of genetic parameters for growth traits in cultured clam *Meretrix meretrix* (Bivalvia: Veneridae) using the Bayesian method based on Gibbs sampling. Aquaculture Research **42**, 240–247.
- Wang, Z., Weber, J. L., Zhong, G., and Tanksley, S. D. (1994) Survey of plant short tandem repeats. Theoretical and Applied Genetics **88**, 1–6.
- Weber, J. L. (1990) Informativeness of human (dC–dA) n(dG–dT)n polymorphisms. Genomics **7**, 524–530.
- Whan, V. A., Wilson, K. J., and Moore, S. S. (2000) Two polymorphic microsatellite markers from novel *Penaeus monodon* ESTs. Animal Genetics **31**, 143–144.
- Zeng, S., Xiao, G., Guo, J., Fei, Z., Xu, Y., Roe, B. A., and Wang, Y. (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. BMC genomics **11**: 94, doi:10.1186/1471-2164-11-94.