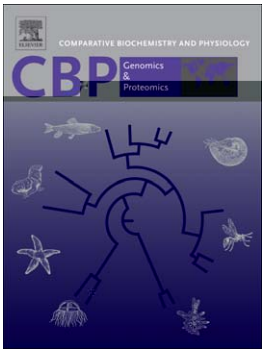


Complete mitochondrial genome of *Anadara vellicata* (Bivalvia: Arcidae): A unique gene order and large atypical non-coding region

Shao'e Sun, Lingfeng Kong, Hong Yu, Qi Li

PII: S1744-117X(15)00054-4
DOI: doi: 10.1016/j.cbd.2015.08.001
Reference: CBD 374

To appear in: *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*

Received date: 20 April 2015
Revised date: 4 August 2015
Accepted date: 17 August 2015

Please cite this article as: Sun, Shao'e, Kong, Lingfeng, Yu, Hong, Li, Qi, Complete mitochondrial genome of *Anadara vellicata* (Bivalvia: Arcidae): A unique gene order and large atypical non-coding region, *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics* (2015), doi: 10.1016/j.cbd.2015.08.001

Complete mitochondrial genome of *Anadara vellicata* (Bivalvia: Arcidae): a unique gene order and large atypical non-coding region

* * * (

Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao 266003, China

Running title: Complete mitochondrial genome of *Anadara vellicata*

ms. has 35 pages, 6 figures, 2 tables, 2 suppl. figures, 2 suppl. tables

*Corresponding author: Tel: +8653282031622. Fax: +8653282032773.

E-mail: qili66@ouc.edu.cn

Abstract

The mitochondrial (mt) genome is a significant tool for investigating the evolutionary history of metazoan animals. The family Arcidae belongs to the superfamily Arcacea in the bivalve order Arcoida, comprising about 260 species. Currently, three complete mitochondrial genomes are available in GenBank, representing 1 subfamilies and 2 genera. Here we present the complete mitochondrial genome sequence of *Anadara vellicata* (Bivalvia: Arcidae), the first report of complete mitogenome from *Anadara,* Arcidae, and compared its sequence with other available Arcidae mitogenomes. The *A. vellicata* mitogenome is 34,147 bp in length, including 12 protein-coding genes (PCGs), 25 transfer RNAs (tRNAs), 2 ribosomal RNA (tRNA) genes and non-coding regions (NCR) (20,722bp). The nucleotide composition of the genome is A+T biased, accounting for 61.03%, with negative AT skew (-0.12) and positive GC skew (0.41). We report the evidence of alloacceptor tRNA gene recruitment (*trnY-trnL2*). A conserved 23 bp-long sequences was 1 *rrnS*. Most of the non-coding sequences (16,112bp) are observed within one segment. In the NCR, the tandem repeat (TR) region is 1,143 bp, comprising six tandem repeats with 189 bp to 192 bp in length. In addition, a long thymine-nucleotide stretch (T-stretch) was detected in the NCR of *A. vellicata*. The gene order and transcriptional polarity of the protein-coding genes is identical to other Arcidae species. tRNA genes are rearranged, making the gene order unique. The results support that mt gene arrangement among Arcidae species is not random, but correlated with their evolutionary relationships.

Abbreviations: Sb, *Scapharca broughtonii*; Sk, *Scapharca kagoshimensis*; Tg, *Tegillarca granosa*; Av, *Anadara vellicata*; *atp6*, ATPase subunit 6 genes; *Cytb*, cytochrome b gene; *cox1-3*, cytochrome *c* oxidase subunitsI-III genes; NCR, non-coding region; *nad1-6* and *nad4l*, NADH dehydrogenase subunits 1-6 and 4L genes; rRNA , ribosomal RNA; *rrnL* and *rrnS*, large and small subunits of ribosomal RNA genes; tRNA, transfer RNA; PCG, protein coding gene; TR, tandem repeat; mtDNA, Mitochondrial DNA; mt, mitochondrial; NJ, Neighbor-joining; ML, maximum likelihood; SDM, strand displacement model; TR, tandem repeat; O_R, replication origin

Key words: *Anadara vellicata*, mitochondrial genome, gene order, non-coding region

1. Introduction

Most metazoan mitochondrial genomes are covalently closed circular molecules which range in size from 14 to 42 kb (Wolstenholme, 1992). However, as more animal mtDNA sequences are analyzed and documented, many newly analyzed mtDNAs are much larger than 14-42 kb. For example, the mt genomes of two Arcidae species, *Scapharca broughtonii* (Liu et al., 2013) and *S. kagoshimensis* (Sun et al., 2014), deviate from the size of typical metazoan mtDNAs, with a length of 46,985 bp and 46,713 bp, respectively. The mt genome typically encodes 37 genes: 13 protein-coding genes (*atp6*, *atp8*, *cox1-3*, *Cytb*, *nad1-6* and *nad4l*), two ribosomal RNAs (*rrnS* and *rrnL*), 22 transfer RNAs genes and a large NCR. This last is often called a control region because it contains sequences essential for the initiation of transcription and replication of the mitogenome (Shadel and Clayton, 1997). In some mitogenomes, such as most Unionidean bivalves, a few intergenic nucleotides have been found, without the large non-coding sequences (Soroka, 2010). Mitochondrial DNA has been extensively used in phylogenetic analyses due to its lack of recombination, maternal inheritance and a higher rate of base substitution than nuclear genes (Brown et al., 1979; Gissi et al., 2008; Krabayashi et al., 2008). Particularly, complete mtDNA sequences can be informative at deep phylogenetic levels (Curole and Kocher, 1999) and their phylogenetic utility has been demonstrated in various animal taxa, including invertebrates (Yokobori et al., 2007; Carapelli et al., 2007; Doucet-Beaupré et al., 2010).

Bivalves are the second largest groups of molluscs after the gastropods, and they are important members of the marine and freshwater ecosystems, including commercially important shellfish in aquaculture or wild harvest (Bieler and Mikkelsen, 2006). To date, at least 90 bivalve mitogenomes from ten orders of Bivalvia have been deposited in public databases (NCBI). Compared with other metazoans, bivalves display much variation in terms of mt genome size, number of tRNA genes, and gene arrangement (Gissi et al., 2008). In family Pectinidae for example, mitogenomes exhibit the most variation in genome organization within the Bivalvia, characterized by several unusual features: a high level of variation in tRNA gene number; extensive translocation of genes; and genome size differences caused by highly variable lengths of NCRs (Ren et al., 2010a; Smith and Snyder, 2007). Gene arrangement has been shown to be very powerful characters for reconstructing evolutionary relationships (Yuan et al., 2012a; Ren et al., 2010b). Furthermore, in many

bivalve mitogenomes, the NCR often contains some peculiar patterns (e.g. AT-rich segment, stem-loop structure, tandem repeat array, microsatellite-like element) (Milbury and Gaffney, 2005; Wang et al., 2011; Yuan et al., 2012b; Liu et al., 2013a; Sun et al., 2015). These characteristics have been used to investigate mitogenomic div

A live individual of *A. vellicata* was collected from the coastal water of Guangxi Province, China. The verification of sample identification was made by experts. Total genomic DNA of *A. vellicata* was extracted from adductor muscle by a modification of standard phenol-chloroform procedure as described by Li et al. (2002) and visualized on 1.0% agarose gel.

2.2. Determination of partial sequences

The short fragment of *cox1* was amplified by PCR with primers LCO-1490/HCO-2198 (Folmer et al., 1994). Another short fragment, *rrnS*, was obtained from NCBI data base (GenBank accession no. JN974641).

2.3. Construction of BD GenomeWalker DNA libraries, PCR and DNA sequencing

The BD GenomeWalker DNA libraries were constructed using the BD GenomeWalker Universal Kit (BD Biosciences, San Jose, CA, USA
protocols.

The complete mt genome of *A. vellicata* was amplified using genome walking and subsequently by long PCR. The genome walking protocol involves two nested PCR reactions with a touch-down program that was modified from the BD GenomeWalker Universal Kit User Manual. The fragments of *cox1* and *rrnS* were used to design the initial sets of gene-specific primers, one (GSP1) for original PCR and one (GSP2) for nested PCR, which were used for genome-walking to amplify the mitogenome of *A. vellicata.* The primer sequences used for genome-walking are presented in Supplementary Table 1.

A 3. s containing 2 U Taq DNA polymerase (TaKaRa, Dalian, China), about 100 ng template DNA, 1 M forward and reverse primers, 200 M of each dNTP, 1×PCR buffer and 2 mM $MgCl_2$. The original PCR uses the outer adaptor primer1 (AP1) and the outer gene-specific primer1 (GSP1). The PCR procedures were as follows: 10 s initial denaturation at 94℃, 7 cycles of 30 s at 94℃, 3 min at 72℃, 32 cycles 30 s at 94 ℃, 3 min at 67℃, and 67℃ for an additional 7 min after the final cycle. A 1- original PCR was diluted in 59 *
for nested PCR amplification. The nested PCR uses the outer adaptor primer2 (AP2) and the outer, gene-specific primer2 (GSP2). The

denaturation at 94℃, 5 cycles of 30 s at 94℃,3 min at 72℃, 25 cycles 30 s at 94℃, 3 min at 67℃, and 67℃ for an additional 7 min after the final cycle. This generally produces a single, major PCR product (100 bp-5000 bp) in at least one of the four libraries, which begins in a kno 3 0 B ,

A Long PCR technique (Cheng et al., 1994) were used to amplify a 3 kb fragment to acquire the rest of the mitogenome sequence. The primer sequences were as follows: LrnIF, 3 -TTATGTTTTGGGAGAGGATTACTGTT-1 G * 3 -ATTTCATGGGGTCTTATCGTCTATTT-1 , The reaction conditions were set as follows: 2U of LA Taq polymerase (TaKaRa, Dalian, China), about 50 ng template DNA, 1 M forward and reverse primers, 200 M of each dNTP, 10×LA PCR buffer II (Mg2+ plus), sterile distilled water up to 50 . PCR procedures for the long fragments were：94℃ for 3 min followed by 35 cycles of 94℃ for 30 s，62℃ for 30 s and 68℃ for 10 min. A final extension step of 72℃ for 10 min was added.

PCR products were purified with EZ-10 spin column DNA gel extraction kit (Sangon Biotech), and then directly sequenced with the primer walking method. The sequencing was conducted on an ABI PRISM 3730 (Applied Biosystems) automatic sequencer in Beijing Genomics Institute (BGI) using standard Sanger sequencing chemistry.

2.4. Gene annotation and sequence analysis

Overlapping fragments obtained from sequencing were assembled with the Seqman program from DNASTAR (http://www.DNASTAR.com). The protein coding genes were analyzed with ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and BLASTx using the invertebrate mitochondrial genetic code. The *atp8* gene was attempt to be detected using HMMER web server (Finn et al., 2011; Eddy, 2011) (as it was not actually detected). The relative synonymous codon usage (RSCU) values of each PCG were calculated using MEGA 5 (Tamura et al., 2011). The positions of tRNA genes were determined by ARWEN (Laslett and Canback, 2008) and DOGMA (Wyman et al., 2004) using the mito/chloroplast or invertebrate genetic code and the default search mode. The rRNA genes were identified by their similarity to published gene sequences and by using BLAST searches (http://www.ncbi.nlm.nih.gov/BLAST/).

The TR sequences were searched by Tandem Repeats Finder 4.0 (Benson, 1999). The

palindrome ,* 0. . 3 , Prediction of potential secondary structure was performed by the online version of the mfold software, version 3.2 (Zuker, 2003), applying default settings. When multiple secondary structures were possible, the most stable (lowest free energy (∆G) was used.

The gene map of the *A. vellicata* mt genome was generated with the program CGView (Stothard and Wishart, 2005). The complete genome sequence is available from NCBI GenBank (GenBank accession no. KP954700).

The ratios of nonsynonymous and synonymous substitutions rates (Ka/Ks) were estimated based on the Maximum -Likelihood (ML) method (Yang and Nielsen, 2000) using KaKs_Calculator 2.0 (Wang et al., 2010a) with the YN model in the twelve PCG of all Arcidae species. The base composition and skewness analyses were performed and compared among all four Arcidae genomes. The A+T content values were computed using Editseq program from DNASTAR. The GC and AT skews described strand bias were calculated according to the formulae by Perna and Kocher (1995), AT skew = (A-T)/(A+T); GC skew = (G-C)/(G+C), where A, T, G and C are the occurrences of the four nucleotides. The skew calculations measured the deviations from A=T and G=C, respectively.

2.5. Gene order comparisons and phylogenetic analysis

Along with mitochondrial genome sequence of *A. vellicata*, all currently available mitochondrial genomes from Arcidae, including *S. broughtonii* (AB729113), *S. kagoshimensis* (KF750628) and *Tegillarca granosa* (KJ607173), were used in comparative gene order and phylogenetic analysis. Mitochondrial gene order rearrangements were analyzed directly by visually comparing proposed rearrangement steps between each mtDNA genome.

The phylogenetic relationships were reconstructed based on nucleotide sequences of 12 PCGs. The root of phylogenetic tree was determined by using *Crassostrea gigas* (AF177226) and *Crassostrea hongkongensis* (EU266073) as the outgroups. Each gene were aligned with MAFFT (Katoh et al., 2005) based on their nucleotide sequences using default settings. The final nucleotide sequences of each gene were then concatenated into single contigs (7160 bp) for phylogenetic analyses. The best-fit nucleotide substitution models for each data partitions were selected by jModelTest (Posada, 2008). Phylogenetic trees were built by ML analysis

using RAxML Black-Box webserver (http://phylobench.vital-it.ch/raxml-bb/index.php; Stamatakis et al., 2008) with GTR + G substitution model to each partition. For ML analysis, 1000 bootstraps were used to estimate the node reliability. A neighbor-joining tree of tRNA genes based on p-distances was constructed using the MEGA 5 program.

3. Results and discussion

3.1 Genome organization of A vellicata

The complete mitogenome sequence of *A. vellicata* is 34,147 bp in length containing 12 protein coding genes (PCG), 2 rRNA genes, and 25 tRNA genes (Fig 1, Table 1). All 39 genes were encoded on the (+) strand. Only one overlap between adjacent genes was found in *A. vellicata*, 4 bp between *cox1* and *nad5*. The 4 bp *cox1-nad5* overlap was also observed in *S. broughtonii* mitogenome (Liu et al., 2013a). Of the 34,147 bp sequence, 13,425 bp are coding DNA, and 20,722 bp are non-coding DNA. The non-coding DNA is distributed throughout the genome in regions ranging in length from 1 to 5058 bp (Fig 1, Table 1).

3.2 Nucleotide composition

The overall base composition of the (+) strand for the mt genome sequence of *A. vellicata* was as follows: A= 26.78%, C = 11.45%, G= 27.52%, T = 34.25%. The A+T bases comprised 61.37% of the PCGs, 58.69% of the rRNAs, 55.10% of the tRNAs and 61.60% of the NCRs, giving a total A+T content of 61.03% in this mt genome. This figure was higher than that in *T. granosa* (60.17%) and lower than *S. broughtonii* (67.89%) and *S. kagoshimensis* (62.75%) (Table 2). For the entire *A. vellicata* mtDNA, the AT and GC skews for the (+) strand were -0.122 and 0.412, respectively, which were significant compared with those of other Arcidae species characterized to date (AT skews ranging from -0.166 to -0.099, and GC skews from 0.358 to 0.447). Nucleotide compositions of the individual gene/region of *S. broughtonii, S. kagoshimensis, T. granosa* and *A. vellicata* were calculated (Table 2). Among different types of genes, protein coding genes and tRNA genes display negative AT skews, whereas rRNA genes (*rrnS* in *S. broughtonii*, *S. kagoshimensis*, and *A. vellicata*, *rrnL* in *S. broughtonii*) show positive AT skews; all types of genes display positive GC skews.

Among the PCGs of *A. vellicata*, the *nad4l* gene has strongly positive GC skew (0.570),

and *Cytb* gene shows weakly positive GC skew (0.240). Replication has be considered as the major source of GC skew variation in the mitogenomes (Sahyoun et al., 2014). Reyes et al. (1998) reported a negative correlation between the GC skew (measured on one strand) and the time that the other strand spends single stranded during replication based on the strand displacement model (SDM) (Clayton, 1991). Sahyoun et al. (2014) believed the GC skew is correlated with the distance from the replication origins, because DNA is in a single stranded state for a time depend on its relative position to the origins. Based on the hypothesis, the GC skew values from this study have provided evidence for the location of the replication origins in Arcidae mitogenomes.

3.3. Protein coding genes

Among the expected 13 protein-coding genes, 12 were identified in *A. vellicata*. No *atp8* coding sequence was detected in this process. Absence of this gene has been suggested for most bivalve species with the exceptions of the venerid *Venerupis philippinarum*, the hiatellid *Hiatella arctica* and Unionoida species (Breton et al., 2010). The *atp8* gene has also been found in *Meretrix lusoria* (Wang et al., 2010b), *Mimachlamys senatoria* and *M. nobilis* (Wu et al., 2013). It was proposed that because the *atp8* protein is characterized by short and variable length, and by high variability in amino acid composition, its absence in some species could be the consequence of annotation difficulties (Gissi et al., 2008). Compared with the available mitogenomes of Arcidae species, the complete mtDNA of *A. vellicata* has the shortest *cox1* gene (1455 bp), and *nad1* gene (570 bp).

Excluding stop codons, the mitogenome of *A. vellicata* encodes 3240 amino acids. The four most predominant codon families are Phe, Leu (UUR), Val, and Gly. The relative synonymous codon usage (RSCU) also reflected the nucleotide composition bias (Supplementary Fig 1). The four- and two-fold degenerate codon usage was A + T biased in the third codon positions. In *A. vellicata* mitogenome, it was notable that, for Phe (UUY), the RSCU was 1.86 for UUU and only 0.14 for UUC. Similarly, for Leu (UUR), the RSCU was 2.63 for UUA and 1.62 for UUG.

The estimation of nonsynonymous (Ka) and synonymous (Ks) substitution rates is quite useful for understanding the selective constraints acting on the protein-coding sequences across closely related species (Ohta, 1995; Fay and Wu, 2003). In order to detect the influence

of selection pressure in Arcidae species, the numbers of Ka, Ks and their ratios were calculated for all pairwise comparisons among the four Arcidae (Supplementary Table 2). The ratio of Ka/Ks in all 12 protein-coding genes varied from 0.0019 for *Cytb* in *S. broughtonii* and *A. vellicata* to 0.9913 for *nad2* in *S. kagoshimensis* and *A. vellicata*, supporting the existence of different mutation constraints among genes. Most of the nonsynonymous substitutions are localized in the *NADH* dehydrogenase complex genes, suggesting a relaxation of purifying selection in the *NADH* complex genes in Arcidae mitogenomes compared with the cytochrome *c* oxidase subunit (*cox1-cox3*) genes and cytochrome *b*. An alternative explanation given the very high ratio in *nad2* would be positive selection. Although all ratios less than 1 is consistent with purifying selection/varying constraints, ratios close to 1 are unusual for mt genes, and may suggest positive selection.

3.4 Transfer RNA (tRNA) genes

The mitogenome of *A. vellicata* contained 25 tRNA genes, which can be folded into expected cloverleaf secondary structures with normal base pairings. All the tRNAs are encoded on the (+) strand, ranging in size from 63 bp (*trnS*) to 80 bp (*trnN* (*AAU*))with

B A. Four tRNA

genes are present more than once: there are 3 copies of *trnK* gene, 2 copies of *trnF* gene, 2 copies of *trnL* gene, and 2 copies of *trnN* gene. It seems that tRNA gene multiplication occurred frequently during the evolution of bivalve mtgenomes (Wu et al., 2012). For example, two *trnD* genes were found in scallop *Mizuhopecten yessoensis* (Wu et al., 2009). The *trnQ* gene duplication or multiplication events have been observed in ten species from five different families (Wu et al.,2012). The tRNA genes (*trnM*, *trnK*, *trnL*, *trnF*, *trnE*, *trnS*, *trnN*, *trnC*, *trnI*, *trnY*, *trnR*) appeared at least two copies in the mitogenomes of other Arcidae species mitogenomes (Liu et al., 2013a; Sun et al., 2014; Sun et al., 2015). The presence of two *trnM* genes has been treated as a common feature of bivalve mt genomes (Xu et al., 2012; Yu and Li, 2012).

The traditional view of tRNA evolution presumes that alloacceptor tRNAs coevolve with the genetic code while isoacceptor tRNA genes evolve by gene duplication from a common ancestor (Wong, 1975; Xue et al., 2003). However, several studies reported that at least some tRNAs could have evolved independently of the genetic code by changing both the anticodon

and acceptor identities of duplicated alloacceptor tRNA genes (Burger et al., 1995; Cedergren ,* . , ,* * been observed in the mt genomes of several organisms (Lavrov and Lang, 2005; Wang and Lavrov, 2011).

In our study, phylogenetic analyses reveals that the *trnL2* is most closely related to *trnY* (Fig. 2A), and they share 63.0% sequence identity (excluding the loop regions) (Fig. 2B). A possible parsimonious explanation for the existence of *trnL2* in the mt genome of *A. vellicata* is that this tRNA gene was derived from a recently duplicated *trnY* gene via an alloacceptor tRNA gene recruitment process, changing the acceptor identity and anticodon sequence. The *trnK1* and *trnK2* show high sequence similarity (98.6%), strongly suggesting that the second *trnK* gene may be derived by gene duplication recently (Fig. 2B).

3.5 Ribosomal RNA (rRNA) genes

Alignment of the Arcidae *rrnS* genes 1 *suggesting the inconsistent annotations. In *A. vellicata* 1 boundary of *rrnS* identification, a 23 bp-long sequence A A AA AA was used, which was folded into a stem-loop structure by mfold software. The stem-loop is formed by nucleotides 16,820 bp to 16842 bp, leaving a 1-nucleotide tail (Fig 3). Stem-loop structure was also used in the identification of *rrnS* in other species of Bivalvia, such as *Crassostrea virginica* (Milbury and Gaffney, 2005), *M. lusoria* (Wang et al., 2010b), and *M. lamarckii* (Wang et al., 2011). And in some Veneroida species, *V. philippinarum*, *M. petechialis*, *Acanthocardia tuberculata*, and *Loripes lacteus*, this structure 1 *rrnS* gene (Wang et al., 2010b). Thus, we inferred that the *rrnS* gene of *A. vellicata* is 673 bp (16,171-16,843) long, with an A+T content of 52.45%.

In order to re-annotate the boundary of *rrnS* in Arcidae mitogenomes, the *rrnS* genes and their sequences 1 aligned. The 23 bp-long sequences were detected 1 of *rrnS* genes in other three Arcidae mitogenomes, with highly conserved stem-loop structures, which were similar to the structure found in *A. vellicata* (Fig 3). The strong similarit 1 *rrnS* in Arcidae mitogenome suggested that the more accurate putative lengths of *rrnS* are 704 bp (25,857-26,560) in *S. broughtonii*, 706 bp (26,022-26,727) in *S. kagoshimensis*, and 677 bp (17,577-18,253) in *T. granosa*.

Hence, at least in Arcidae mitogenomes, this conserved 23 bp-long sequences may be used as

1 *rrnS*.

Identification of the large subunit rRNA gene (*rrnL*) in *A. vellicata* was accomplished by comparison with other Arcidae *rrnL* gene. The *rrnL* was 1350 bp in length, located between *trnV* and *trnA*. The length of *rrnL* gene in *A. vellicata* is the largest yet reported in the family Arcidae.

3.6 Non-coding regions

In all, 20,722 bp of the *A. vellicata* mt genome was predicted to be non-coding sequence, accounting for 60.68% of the entire mtDNA. The A+T content of NCR is 61.60%. Most of the non-coding sequences (16,112bp) were observed within one segment, and within this segment all of the sequence, except 2 dispersed tRNAs, which totaled 131 bp, was predicted to be non-coding DNA. The investigation of the non-coding regions of the four Arcidae species revealed distinct structural patterns. Firstly, there is high variability in the lengths of the non-coding regions in the four Arcidae (*A. vellicata*, 20,722 bp, *S. broughtonii*, 31,658 bp; *S. kagoshimensis*, 32,982 bp; *T. granosa*, 16,394 bp), and alignments of these sequences show very low identities between each other. The size variation of Arcidae mt genomes is due to the different length of the non-coding regions. Secondly, the distribution of the non-coding sequences was different. Most of the non-coding DNAs were observed within two distinct zones (between *cox2* and *nad6*, *nad2* and *cox1*) in the mt genomes of *S. broughtonii* and *S. kagoshimensis*. However, *A. vellicata* and *T. granosa* has only one large concentrated non-coding region, located between *nad2* and *cox1*.

In the large concentrated NCR of *A. vellicata*, the TR region is 1,143 bp (position 31,003-32,145), comprising six DNA repeats with 189 bp to 192 bp in length . Each of the tandem repeats included an inverted repeats separated by seven nucleotides (Fig 4A). All the six inverted repeats and the separating nucleotides could fold into stem-loop secondary structures with several mismatches in the stems (Fig. 4B). The sequence identity of the repeat units in *A. vellicata* reached 91.1%.

The tandem repeat region was also found in other Arcidae mitogenomes, but both the length and copy number of the repeat units were different. The largest repeat region (321 bp) in *S. broughtonii* consisted of thirty-one tandem repeat units (19 bp for each). The largest

repeat region (541 bp) of *S. kagoshimensis* was comprised of nine complete tandem repeats units, which were 60 to 61 bp in length. In *T. granosa*, five repeat units, ranging from 124 to 132 bp and a 77 bp partial repeat unit were detected in the largest repeat region (719 bp). Although these repeat unit sequences lack identities among Arcidae, all of them could be folded into stem-and-loop secondary structures (Supplementary Fig 2). The occurrence of tandem repeats is explained by different models, but the mtDNA replication has been suggested to be the primary mechanism causing gain or loss of repeats through slippage-strand mispairing (Moritz and Brown, 1987; Levinson and Gutman, 1987; Broughton and Dowling, 1994). Moreover, potential stem-loop structures in a repeated unit and its flanking part have been demonstrated to cause an increase in slipped-strand mispairing frequency (Boore, 2000; Levinson and Gutman, 1987).

The NCR of *A. vellicata* contains a microsatellite-like $(AT)_{18}$ element at positions 23,545-23583. The similar $(AT)_n$ region, simple sequence repeat regions, has also been reported in mitogenome of *Crassostrea virginica* (Milbury and Gaffney, 2005) and *Solen grandis* (Yuan et al., 2012b).

3.7 Recognition sequences of replication origins

A 23 bp poly-T stretch (23,708-23,730) was found in the NCR of *A. vellicata*. The T-stretches were also revealed in the NCR of other available Arc3(t)6(i)51 017 TtandeearriSo

is involved in the recognition of the O_R in insect mtDNA (Saito et al., 2005). Based on this hypothesis, the long T-stretches could also play a crucial role in the replication initiation of Arcidae mtDNA.

3.8 Unique gene order and phylogenetic implication

In this study, we focus on the gene rearrangement of *A. vellicata* in the family Arcidae.

long non-coding sequences (20,772 bp). Most of the non-coding regions (16,112bp) were concentrated in one segment. The AT content of the (+) strand is 61.03%, with a negative AT skew (-0.12).

Most of the nonsynonymous substitutions are localized in the *NADH* dehydrogenase complex genes, indicating that these genes bear less selective pressure compared with other mitochondrial protein-coding genes. The ratio in *nad2* close to 1 would be positive selection.The analysis suggested that *trnL2* was derived from a recently duplicated *trnY* gene via an alloacceptor tRNA gene recruitment process, whereas the *trnK2* gene may be derived by gene duplication. A conserved 23 bp-long sequences was 1 terminus of *rrnS*.

Within the NCR, a TR region was found, which is the largest set of tandem repeat found in the mitochondrial genomes of Arcidae. Each of the tandem repeat motif formed five or six stem-loop structures when the sequence is folded to minimize the free energy of the structure. Moreover, we found a microsatellite-like $(AT)_{18}$ element and a long T-stretch in the mtDNA of *A. vellicata*. The T-stretch participated in the formation of or be positioned adjacent to possible stem-loop structures, which could involved in the recognition of the O_R in *A. vellicata* mtDNA.

The gene order and transcriptional polarity of the PCGs is identical to other Arcidaes. Nevertheless, tRNA genes rearrangement, making the gene arrangement unique. The studies based on the conbination of gene order and phylogenetic analysis suggest that gene arrangement among Arcidae species is correlated with their evolutionary relationships. The distinct mitogenome architecture revealed by *A. vellicata* has advanced our understanding of Arcidae mitogenomic diversities and evolution.

Acknowledgments

This study was supported by research grants from National Marine Public Welfare Research Program (201305005), National Natural Science Foundation of China (41276138), and Doctoral Program of Ministry of Education of China (20130132110009).

References

Anderson, S., De Bruijn, M.H.L., Coulson, A.R., Eperon, I.C., Sanger, F., Young, I.G., 1982.

W.R., 2010. Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. BMC Evol. Biol. 10, 50.

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLoS Comp. Biol. 7, e1002195.

Fay, J.C., Wu, C.I., 2003. Sequence divergence, functional constraint, and selection in protein evolution. Annu. Rev. Genomics. Hum. Genet. 4, 213-235.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3, 294-299.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic. Acids. Res. 39, 29-37.

Gissi, C., Iannelli, F., Pesole, G., 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity, 101, 301-320.

Hixson, J.E., Wong, T.W., Clayton, D.A., 1986. Both the conserved stem-loop and divergent 5' -flanking sequences are required for initiation at the hum an mitochondrial origin of light-strand DNA replication. J. Biol. Chem. 261, 2384-2390.

Katoh, K., Kuma K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic. Acids. Res. 33, 511-518.

Kornberg, A., Baker, T.A., 1992. DNA Replication. Ed. 2, pp. 1-52. WH. Freeman, New York.

Krabayashi, A., Sumida, M., Yonekawa, H., Glaw, F., Vences, M., Hasegawak, M., 2008. Phylogeny, recombination, and mechanisms of stepwise mitochondrial genome reorganization in mantellid frogs from Madagascar. Mol. Biol. Evol. 25, 874-891.

Laslett, D., Canback, B., 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24, 172-175.

Lavrov, D.V., Lang, B.F., 2005. Transfer RNA gene recruitment in mitochondrial DNA. Trends. Genet. 21, 129-133.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203-221.

Li, Q., Park, C., Kijima, A., 2002. Isolation and characterization of microsatellite loci in the Pacific abalone, *Haliotis discus hannai*. J. Shellfish. Res. 21, 811-815.

* ,* * G ,* * ,* * G* * , 0. . 3,

yeast Saccharomyces cerevisiae genome. Curr. Genet. 47, 289-297.

Liu, Y.G., Kurokawa, T., Tanabe, T., Watanabe, K., 2013a. Complete mitochondrial DNA sequence of the ark shell *Scapharca broughtonii*: an ultra-large metazoan mitochondrial genome. Comp. Biochem. Physiol. D: Genomics Proteomics. 8 ,72-81.

Liu, Y.G., Kurokawa, T., Sekino, M., Tanabe, T., Watanabe, K., 2013b. Tandem repeat arrays in the mitochondrial genome as a tool for detecting genetic differences among the ark shell *Scapharca broughtonii*. Mar. Ecol. 35, 273-280.

Milbury, C.A., Gaffney, P.M., 2005. Complete mitochondrial DNA sequence of the eastern oyster *Crassostrea virginica*. Mar. Biotechnol. 7, 697-712.

Moritz, C., Brown, W.M., 1987. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. Proc. Natl. Acad. Sci. USA. 84, 7183.

Ohta, T., 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40, 56-63.

Oliver, P.G., Holmes, A.M., 2006. The Arcoidea (Mollusca: Bivalvia): a review of the current phenetic-based systematics. Zool. J. Linn. Soc. 148, 237-251.

Perna, N.T., Kocher, T.D., 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. 41, 353-358.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Molecular biology and evolution. 25, 1253-1256.

Ren, J., Shen, X., Jiang, F., Liu, B., 2010a. The mitochondrial genomes of two scallops, *Argopecten irradians* and *Chlamys farrei* (Mrollusca: Bivalvia): the most highly rearranged gene order in the family Pectinidae. J. Mol. Evol. 70, 57-68.

Ren, J., Liu, X., Jiang, F., Guo, X., Liu, B., 2010b. Unusual conservation of mitochondrial gene order in *Crassostrea* oysters: evidence for recent speciation in Asia. BMC Evol. Biol. 10, 394.

Reyes, A., Gissi, C., Pesole, G., Saccone, C., 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15, 957-966.

Sahyoun, A. H., Bernt, M., Stadler, P. F., Tout, K. 2014. GC skew and mitochondrial origins of replication. Mitochondrion, 17, 56-66.

Saito, S., Tamura, K., Aotsuka, T., 2005. Replication origin of mitochondrial DNA in insects.

Genetics, 171, 1695-1705.

Saks, M.E., Sampson, J.R., Abelson, J., 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. Science. 279, 1665-1670.

Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. 66, 409-435.

Sinden , R.R., 1994. DNA Structure and Function, pp. 58-94. Academic Press, San Diego.

Smith, D.R., Snyder, M., 2007. Complete mitochondrial DNA sequence of the scallop *Placopecten magellanicus*: evidence of transposition leading to an uncharacteristically large mitochondrial genome. J. Mol. Evol. 65, 380-391.

Soroka, M. 2010. Characteristics of mitochondrial DNA of unionid bivalves (Mollusca: Bivalvia: Unionidae). II. Comparison of complete sequences of maternally inherited mitochondrial genomes of *Sinanodonta woodiana* and *Unio pictorum*. Folia Malacologica, 18, 189-209.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758-771.

Stothard, P., Wishart, D.S., 2005. Circular genome visualization and exploration using CGView. Bioinform. 21, 537-539.

Sun, S.E., Kong, L.F., Yu, H., Li, Q., 2014. The complete mitochondrial genome of *Scapharca kagoshimensis* (Bivalvia: Arcidae). Mitochondrial DNA, 1-2.

Sun, S.E., Kong, L.F., Yu, H., Li, Q., 2015. The complete mitochondrial DNA of *Tegillarca granosa* and comparative mitogenomic analyses of three Arcidae species. Gene, 557, 61-70.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731-2739.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., Yu, J., 2010a. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics, Proteomics & Bioinformatics. 8, 77-80.

Wang, H., Zhang, S., Li, Y., Liu, B., 2010b. Complete mtDNA of *Meretrix lusoria* (Bivalvia: Veneridae) reveals the presence of an *atp8* gene, length variation and heteroplasmy in the control region. Comp. Biochem. Physiol. D: Genomics Proteomics. 8, 72-81.

Wang, H., Zhang, S., Xiao, G., Liu, B., 2011. Complete mtDNA of the *Meretrix lamarckii* (Bivalvia: Veneridae) and molecular identification of suspected *M. lamarckii* based on the whole mitochondrial genome. Mar. Genom. 4, 263-271.

Wang, X., Lavrov, D.V., 2011. Gene recruitment-a common mechanism in the evolution of transfer RNA gene families. Gene. 475, 22-29.

Wu, X., Xu, X., Yu, Z., Kong, X., 2009. Comparative mitogenomic analyses of three scallops (Bivalvia: Pectinidae) reveal high level variation of genomic organization and a diversity of transfer RNA gene sets. BMC Res. Notes. 2, 69.

Wu, X., Li, X., Li, L., Yu, Z., 2012. A unique tRNA gene family and a novel, highly expressed ORF in the mitochondrial genome of the silver-lip pearl oyster, *Pinctada maxima* (Bivalvia: Pteriidae). Gene, 510, 22-31.

Wu, X., Li, X., Yu, Z. 2013. The mitochondrial genome of the scallop *Mimachlamys senatoria* (Bivalvia, Pectinidae). Mitochondrial DNA. 26, 242-244.

Wu, X., Xiao, S., Li, X., Li, L., Shi, W., Yu, Z., 2014. Evolution of the tRNA gene family in mitochondrial genomes of five *Meretrix* clams (Bivalvia, Veneridae). Gene, 533, 439-446.

Wolstenholme, D.R., 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. 141, 173-216.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl Acad. Sci. USA 72, 1909-1912.

Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinform. 20, 3252-3255.

Xu, X.D., Wu, X.Y., Yu, Z.N., 2012. Comparative studies of the complete mitochondrial genomes of four *Paphia* clams and reconsideration of subgenus *Neotapes* (Bivalvia: Veneridae). Gene, 494, 17-23.

Xue, H., Tong, K.L., Marck, C., Grosjean, H., Wong, J.T., 2003. Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. Gene. 310, 59-66.

Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32-43.

Yokobori S, Lindsay D J, Yoshida M, Tsuchiya K, Yamagishi A, Maruyama T, Oshima T

2007. Mitochondrial genome structure and evolution in the living fossil vampire squid, *Vampyroteuthis infernalis*, and extant cephalopods. Mol. Phylogenet. Evol. 44, 898-910.

Yu, H., Li, Q., 2012. Complete mitochondrial DNA sequence of *Crassostrea nippona*: comparative and phylogenomic studies on seven commercial *Crassostrea* species. Mol. Biol. Rep. 39, 999-1009.

Yuan, Y., Li, Q., Yu, H., Kong, L., 2012a. The complete mitochondrial genomes of six heterodont bivalves (Tellinoidea and Solenoidea): variable gene arrangements and phylogenetic implications. PloS one, 7, e32353.

Yuan, Y., Li, Q., Kong, L., Yu, H., 2012b. The complete mitochondrial genome of the grand jackknife clam, *Solen grandis* (Bivalvia: Solenidae): a novel gene order and unusual non-coding region. Mol. Biol. Rep. 39, 1287-1292.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic. Acids. Res. 31, 3406-3415.

Legends

Figure 1. The organization of the mitochondrial genome of *Anadara vellicata*. The full names of protein coding genes, *rrnS* and *rrnL* are listed under Abbreviations. Transfer-RNAs are represented by their one-letter amino acid code. The microsatellite repeat region, T-stretch and tandem repeat (TR) region are indicated on the figure.

Figure 2. (A) Neighbor-joining tree based on uncorrected *p* distances among mitochondrial tRNA genes from *Anadara vellicata*. Portions of the tree discussed in the main text are shown in red. The numbers above branches indicate percentage of bootstrap support based on 1000 replicates (if >50%). (B) Alignment of *trnL2* and *trnY* gene sequence, and that of *trnK1* and *trnK2*. The secondary structure in displayed above the alignment, and the conserved regions are marked with purple asterisk.

Figure 3. Stem- l *rrnS* gene in the four Arcidae species folded by mfold software.

Figure 4. A) Alignment of the six tandem repeats (TRs) of *Anadara vellicata* mitochondrial major non-coding region. For TR 1 the nucleotide sequence is given and for the other repeats dots were representing identical nucleotides. The nucleotides highlighted in red indicate the

conserved inverted repeat in each of the TRs, separated by seven nucleotides, which could be folded into stem-loop structures. B) Stem-loop structure of the inverted repeat and the separating nucleotides, as well as the positions of the six TRs. Variable loop nucleotide is encircled. For each structure the position within the mitogenome is also given.

Figure 5. Potential stem-loop structures adjacent to T-stretches in the family Arcidae. The nucleotides highlighted in red represent the location of the T-stretch.

Figure 6. Phylogenetic relationships among Arcidae derived from Maximum Likelihood (ML) analyses based on partitioned nucleotide sequences of 12 mitochondrial protein-coding genes and gene arrangement map of Arcidae mitochondrial genomes, including tRNA genes. The muber at each node is the bootstrap probability of ML analyses. The scale bar means 0.2 substitutions per site. All genes are transcribed from left-to-right, excluding the non-coding regions. The bars indicate identical gene blocks, but the gene segments are not drawn to scale.

Table 1. Organization of the mitochondrial genome of *Anadara vellicata*.

Gene	Strand	From	To	Size (nts)	Size (aa)	Start codon	Stop codon	Intergenic nucleotides
cox1	+	1	1455	1455	485	ATA	TAG	129
nad5	+	1452	3146	1695	565	ATA	TAA	-4
trnM	+	3159	3225	67				12
nad1	+	3226	3762	537	179	ATG	TAG	0
nad4	+	4321	5496	1176	392	ATG	TAA	558
Cytb	+	7211	8323	1113	371	ATA	TAG	1714
trnF1	+	8336	8400	65				12
cox2	+	8422	9129	708	236	ATA	TAG	21
trnC	+	9198	9264	67				68
nad6	+	9425	9901	477	159	ATG	TAG	160
trnK1	+	9936	10007	72				34
trnK2	+	10486	10557	72				478
atp6	+	10614	11288	675	225	ATG	TAA	56
trnP	+	11323	11389	67				34
trnI	+	11417	11488	72				27
trnG	+	11490	11558	69				1
trnE	+	11560	11630	71				1
trnV	+	11634	11702	69				3
rrnL	+	11714	13063	1350				11
trnA	+	13064	13134	71				0
trnT	+	13142	13212	71				7
trnH	+	13227	13293	67				14
trnQ	+	13330	13398	69				36
nad3	+	13529	13789	261	87	ATG	TAG	130
nad4L	+	13949	14206	258	86	ATG	TAA	159
trnW	+	14212	14279	68				5
trnS	+	14283	14345	63				3
cox3	+	14577	15236	660	220	ATG	TAA	231
trnD	+	15416	15481	66				179
trnL1	+	15648	15714	67				166
trnY	+	15754	15819	66				39
trnN1	+	15831	15900	70				11
rrnS	+	16171	16843	673				270
nad2	+	17014	17685	672	224	ATA	TAA	170
trnR	+	18307	18374	68				621
trnL2	+	18402	18471	70				27
microsatellite repeat region	+	23545	23580	36				5073

			127
			3544
			2055
			1543
			1794

mitochondrial genes of *Anadara vellicata* *oshimensis* (Sk) and *Tegillarca granosa* (Tg).

						(A27878 . r99.2BT1 0

Sb Sk Tg Av Sb

		78	92	48		59	41	27	91	58	45	00	35
rrnS	52.45	53.64	52.74	53.99		**0.071**	**0.045**	**0.014**	-0.034	0.144	0.179	0.179	0.129
rrnL	60.77	63.82	63.74	59.79		-0.031	**0.020**	-0.001	-0.036	0.284	0.274	0.250	0.333
tRNA	55.10	61.60	57.14	56.23		-0.076	-0.110	-0.070	-0.037	0.251	0.190	0.159	0.254
NCR	61.60	68.42	63.15	58.93		-0.059	-0.091	-0.159	-0.006	0.436	0.364	0.121	0.414

Figure 1

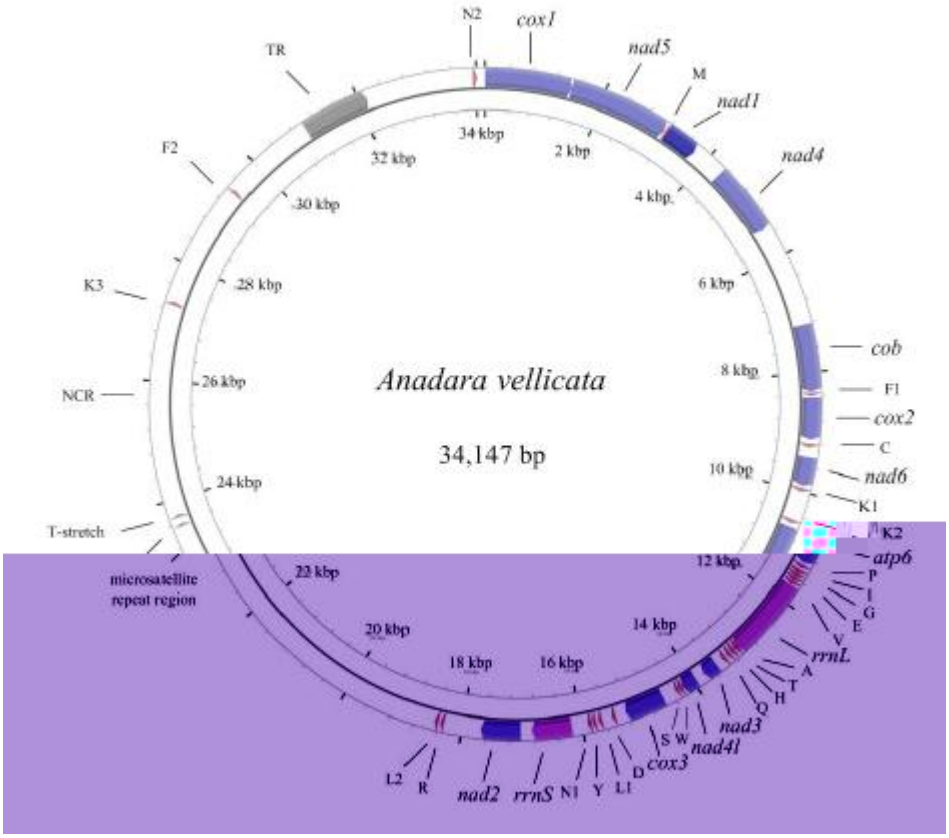

Anadara vellicata
34,147 bp
N2
cox1
nad5
M
nad1
nad4
cob
F1
cox2
C
nad6
K1
K2
atp6
P
I
G
E
V
rrnL
A
T
H
Q
nad3
nad4l
S
W
cox3
D
L1
Y
N1
rrnS
nad2
R
L2
microsatellite
repeat region
T-stretch
NCR
K3
F2
TR
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
24 kbp
26 kbp
28 kbp
30 kbp
32 kbp
34 kbp

Figure 2

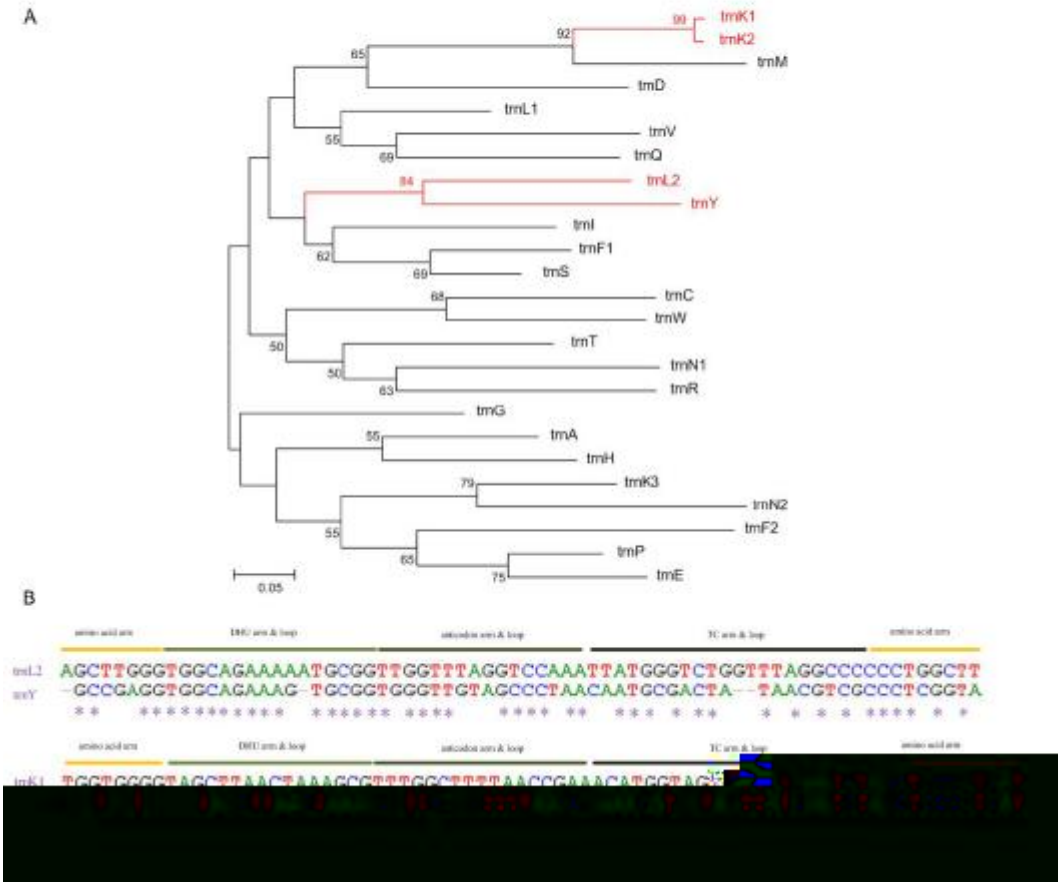

A
trnK1
trnK2
trnM
trnD
trnL1
trnV
trnQ
trnL2
trnY
trnI
trnF1
trnS
trnC
trnW
trnT
trnN1
trnR
trnG
trnA
trnH
trnK3
trnN2
trnF2
trnP
trnE
0.05
B

Figure 3

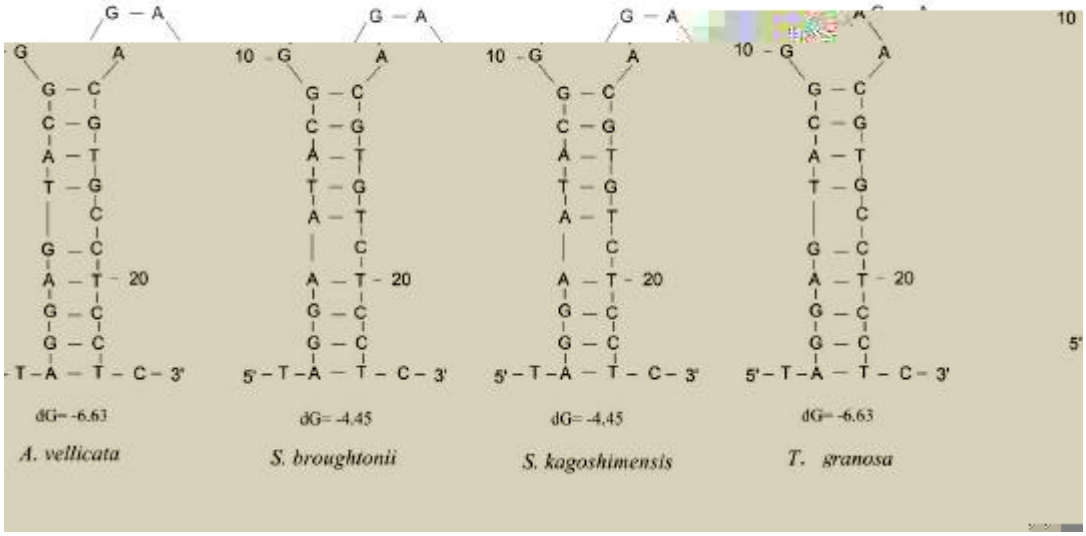

dG= -6.63
A. vellicata
dG= -4.45
S. broughtonii
dG= -4.45
S. kagoshimensis
dG= -6.63
T. granosa

Figure 4

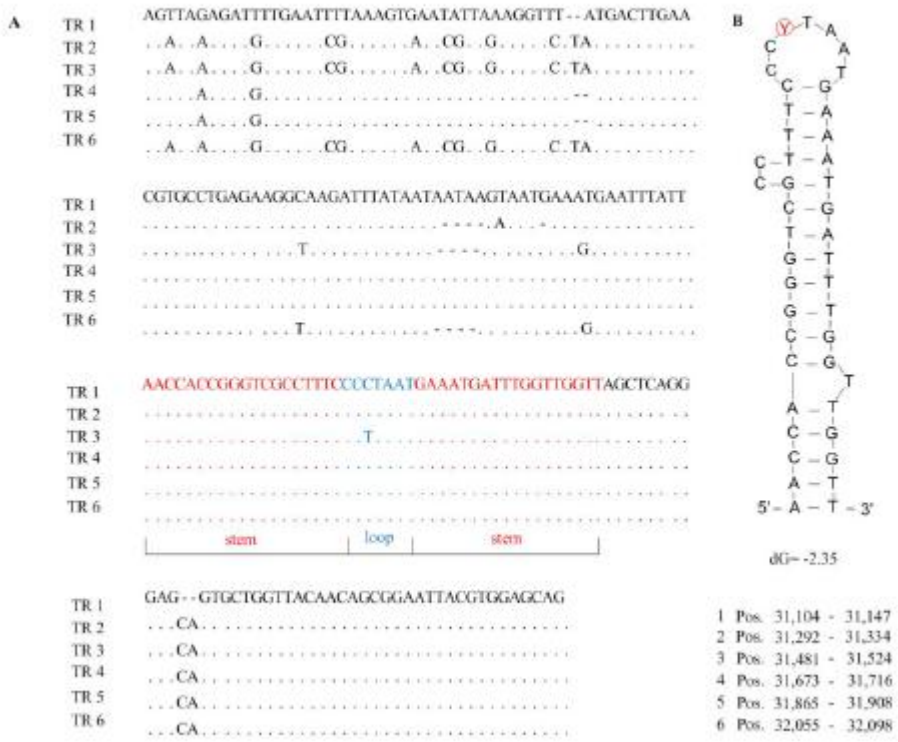

A
TR 1
TR 2
TR 3
TR 4
TR 5
TR 6
AGTTAGAGATTTTGAATTTAAAGTGAATATTAAAGGTTT--ATGACTTGAA
CGTGCCTGAGAAGGCAAGATTTATAATAATAAGTAATGAAATGAATTTATT
AACCACCGGGTCGCCTTTCCCCTAATGAAATGATTTGGTTGGTTAGCTCAGG
stem
loop
stem
GAG--GTGCTGGTTACAACAGCGGAATTACGTGGAGCAG
B
dG= -2.35
1 Pos. 31,104 - 31,147
2 Pos. 31,292 - 31,334
3 Pos. 31,481 - 31,524
4 Pos. 31,673 - 31,716
5 Pos. 31,865 - 31,908
6 Pos. 32,055 - 32,098

Figure 5

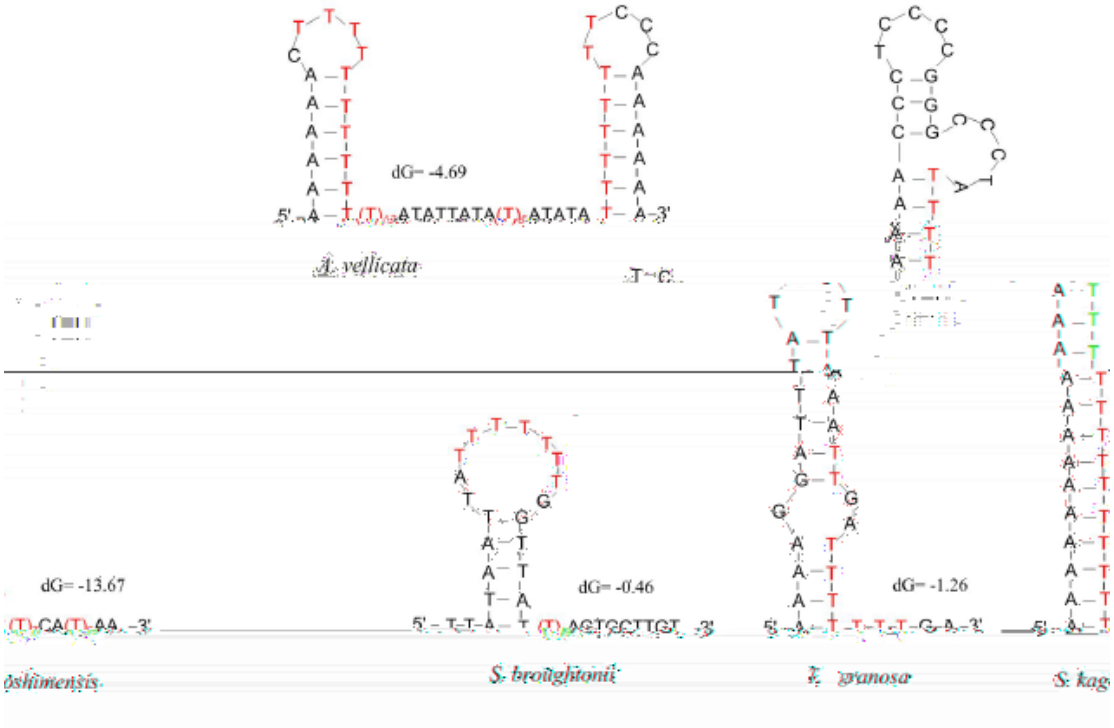
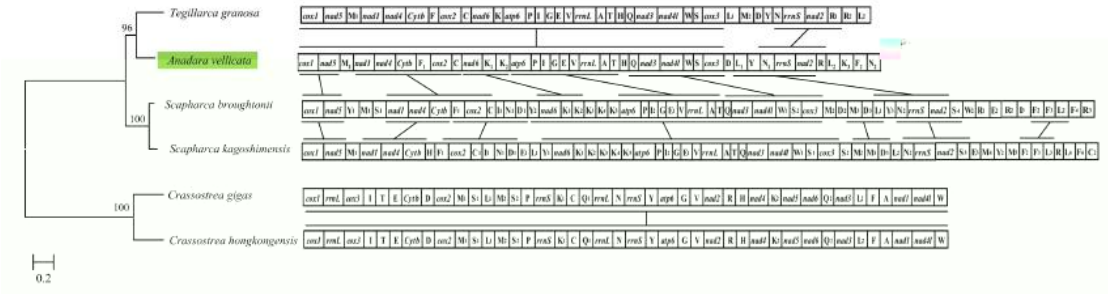


Figure 6


Tegillarca granosa
Anadara vellicata
Scapharca broughtonii
Scapharca kagoshimensis
Crassostrea gigas
Crassostrea hongkongensis
96
100
100
0.2