Molecular Phylogenetics and Evolution 118 (2018) 222–231

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Multiple reversals of strand asymmetry in molluscs mitochondrial genomes, and consequences for phylogenetic inferences

Shao'e Sun[a], Qi Li[a,b,*], Lingfeng Kong[a], Hong Yu[a]

[a] Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao 266003, China
[b] Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, China

ARTICLE INFO

Keywords:
Mitochondrial genome
Strand bias
Mollusca
Phylogenetic analyses

ABSTRACT

Strand asymmetry in nucleotide composition is a remarkable feature of animal mitochondrial genomes. The strand-specific bias in the nucleotide composition of the mtDNA has been known to be highly problematic for phylogenetic analyses. Here, the strand asymmetry was compared across 140 mollusc species and analyzed for a mtDNA fragment including twelve protein-coding genes. The analyses show that almost all species in Gastropoda (except Heterobranchia) and all species in Bivalvia present reversals of strand bias. The skew values on individual genes for all codon positions (P_{123}), third codon positions (P_3), and fourfold redundant third codon positions (P_{4FD}) indicated that CG skews are the best indicators of strand asymmetry. The differences in the patterns of strand asymmetry significantly influenced the amino acid composition of the encoded proteins. These biases are most striking for the amino acids Valine, Cysteine, Asparagine and Threonines, which appear to have evolved asymmetrical exchanges in response to shifts in nucleotide composition. Molluscs with strong variability of genome architectures (ARs) are usually characterized by a reversal of the usual strand bias. Phylogenetic analyses show that reversals of asymmetric mutational constraints have consequences on the phylogenetic inferences, as taxa characterized by reverse strand bias (Heterobranchia and Bivalvia) tend to group together due to long-branch attraction (LBA) artifacts. Neutral Transitions Excluded (NTE) model did not overcome the problem of heterogeneous biases present in molluscs mt genomes, suggested it may not be appropriate for molluscs mt genome data. Further refinement phylogenetic models may help us better understand internal relationships among these diverse organisms.

1. Introduction

The mitochondrial genome (mt genome) of most metazoan animals includes a standard set of 13 protein-coding genes (PCGs), 2 ribosomal RNA (rRNA) genes, 22 transfer RNA (tRNA) genes, and an A + T-rich region. Although there are exceptions, most mitogenomes range in size from 14 to 17 kb. Typically, few intergenic nucleotides exist except for a single large non-coding region, which generally thought to contain elements that control the initiation of replication and transcription of the mitogenome. (Boore, 1999; Lavrov, 2007). Owing to the economized organization, lack of recombination, maternal inheritance (except for Doucet-Beaupré et al., 2010), absence of introns, and higher evolutionary rates, mtDNA sequences are extensively used for comparative and evolutionary genomics, population genetics and phylogenetic inference (Curole and Kocher, 1999; Ballard and Whitlock, 2004; Gissi et al., 2008).

It is known that animal mitochondrial genomes vary significantly in nucleotide composition and almost all show a bias between the two strands of the genome (strand asymmetry) (Perna and Kocher, 1995; Hassanin et al., 2005). In mammals, one strand is G rich, whereas the other strand is G poor, and because they show different buoyant densities in a cesium chloride gradient, the G-rich strand is called "heavy strand" (Anderson et al., 1981). This is different from the plus/minus strand or major/minor coding strand terminology. The plus strand is mostly defined based on the orientation of the *cox1* gene. When most gene are coded on the same strand it is easy to identify this one as the major coding strand (Bernt et al., 2013). Generally, there is more A than T and more C than G on the major or plus strands. However, the strand asymmetry is reversed in some taxa, such as arthropods (Cameron et al., 2007; Hassanin et al., 2005; Kilpert and Podsiadlowski, 2006; Masta et al., 2009; Hassanin, 2006), flatworms (Min and Hickey, 2007), brachiopods (Helfenbein et al., 2001), echinoderms (Scouras and Smith, 2006) and fish (Wang et al., 2007), where A is less than T and C is less than G on the major or plus strands. The underlying mechanism that account for the strong compositional asymmetry observed in mitochondrial genomes has been generally related to replication and

* Corresponding author at: Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao 266003, China.
E-mail address: qili66@ouc.edu.cn (Q. Li).

http://dx.doi.org/10.1016/j.ympev.2017.10.009
Received 29 January 2017; Received in revised form 8 October 2017; Accepted 12 October 2017
Available online 14 October 2017
1055-7903/

transcription processes (Reyes et al., 1998). Because these processes have long been assumed to be asymmetric in the mtDNA and could therefore affect the occurrence of mutations between the two strands (Clayton, 1982; Tanaka and Ozawa, 1994; Reyes et al., 1998).

Sequences of the mt genome have been widely used for addressing phylogenetic questions ranging from population to phylum level (Avise, 2000). With an increasing number of studies the usefulness of mtDNA as a marker for highly divergent lineages was criticized (Curole and Kocher, 1999). Three main characteristics of the mt genome are expected to be problematic for phylogenetic analyses: (1) mutational saturation due to multiple hits is a major source of uncertainty in current molecular phylogeny, and within mt genomes, saturation is all the more important because the mt genomes are more fast-evolving than the nuclear genome (Burger et al., 2003; Moreira and Philippe, 2010); (2) long-branch attraction (LBA) is a very common phenomenon whenever differences of evolutionary rate among different lineages, and the fast-evolving taxa are more prone to group together by chance (false synapomorphies) (Felsenstein, 1978; Moreira and Philippe, 2010); (3) heterogeneity in nucleotide composition among different lineages, such as reversals of strand asymmetry, can mislead phylogenetic inferences because unrelated taxa with similar base compositions rather than genuine phylogenetic signal may be erroneously clustered (Hassanin et al., 2005; Moreira and Philippe, 2010).

Although it is clear that the atypical strand bias has evolved independently among taxa from disparate branches of the tree of life, the patterns of strand bias evolution in molluscs remain unclear. Molluscs exhibit the largest disparity of all animal phyla and rank second behind arthropods in species diversity. (Giribet et al., 2006). Traditionally, the phylum Mollusca is divided into six classes: Gastropoda, Bivalvia, Cephalopoda, Polyplacophora, Scaphopoda and Monoplacophora (Morton and Yonge, 1964). The phylogenetic relationships among the major groups of these major lineages remain one of the most contentious issues in systematics. The usefulness of mt sequences as a marker for reconstructing the phylogeny of molluscs are expected to be problematic, because of the mutational saturation and heterogeneity in nucleotide composition among taxa. The sequence positions have accumulated so many mutations that the present bases or amino acids are essentially random, and therefore contain scant or no evolutive information (Moreira and Philippe, 2010). Additionally, the mt genomes of molluscs are also characterized by a strong compositional bias, which is particularly rich in A and T nucleotides (He et al., 2011; White et al., 2011; Xu et al., 2012). Consequently, the reliability of their phylogenetic relationships was questioned.

In this study, a broad survey of strand asymmetry was analyzed in 140 molluscs mitochondrial genomes. In particular, for each of the 12 protein-coding genes, we investigated the compositional features of all three codon positions P_{123} and third positions of fourfold degenerate codons (P_{4fd}), as well as the effects of DNA asymmetric strand bias on amino acid composition. The sequences were examined to identify the relationship between genome architecture and strand asymmetry in each species. By using the same data matrix, we studied the pattern of molecular evolution and evaluated the effect of strand-bias on phylogenetic inferences. We aim to perform a series of phylogenetic analyses to explore if the Neutral Transitions Excluded (NTE) model is appropriate for molluscs mt genome data to limit this specific problem in tree reconstruction under the Bayesian approach. Our analyses focus on molluscs, but our methods are applicable to any mitochondrial genomes that display reversals of strand-compositional bias.

2. Materials and methods

2.1. Taxonomic sampling

One hundred and forty molluscan mitochondrial genomes were selected for strand asymmetry analyses, represented by three classes (Gastropoda, Bivalvia and Cephalopoda) of molluscs (Supplementary Table 1). Sequences of whole mitochondrial genome sequences were downloaded from GenBank, and the individual mitochondrial protein-coding genes were extracted from each mt genome. The *atp8* gene was excluded.

2.2. Calculation of skew values

The GC and AT skews, which indicate compositional differences between the two strands, were calculated according to the formulae by Perna and Kocher (1995): AT skew = (A − T)/(A + T); GC skew = (G − C)/(G + C). For each mitochondrial genome, AT and GC skews were carried out on all the plus strand, all codon positions (P_{123}), third codon positions (P_3) and fourfold redundant third codon positions (P_{4fd}) of protein-coding genes.

The amino acid compositions were predicted based on the invertebrates mitochondrial genetic code by partitioning the mitochondrial codons into CA-rich codons (CA, AC, CC, AA codons at the first two codon positions), GT-rich codons (GT, TG, GG, TT codons at the first two codon positions), and other codons. Synonymous codon usage of mitochondrial coding sequences was measured by the nucleotide content of G+T or C+A at the third codon positions of fourfold degenerate codon families: GGN (G), GTN (V), CGN (A), ACN (T), GCN (A), and CCN (P). Leucine (L) and Serine (S) were not included in these latter calculations, due to their greater (8-fold) degeneracy. All statistical analysis was performed by using IBM SPSS Statistics 19.

The variability in mt genome AR for every species was estimated by the "AR rate", according to the formulae $(N_{AR} - 1)/(N_{mtDNA} - 1) \times 100$ (e.t. Gissi et al., 2008), where N_{AR} and N_{mtDNA} are the number of different ARs and the number of completely sequenced mitochondril DNAs of that taxa, respectively. Thus, a higher AR rate means more mtDNAs have a different AR.

2.3. Phylogenetic analyses

The twelve-partitioned nucleotide sequences of protein coding genes were aligned with MAFFT (Katoh et al., 2005). Areas of dubious alignment were isolated using Gblocks (Castresana, 2000; Talavera and Castresana, 2007) (default setting) and excluded from the analyses. The best-fit nucleotide substitution models for each data partitions were selected by jModelTest (Posada, 2008), by using the Akaike Information Criterion (AIC). The information of alignment length and DNA substitution models applied to each partition were listed in Supplementary Table 2. In addition, one Annelida species were selected as outgroup: *Platynereis dumerilii* (AF178678).

The phylogenetic relationship was built by three methods: maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). MP analyses were carried out with PAUP 4.0b 10 (Swofford, 2003). Bootstrap proportions (BP_{MP}) were obtained from 1000 replicates by using 10 replicates of random stepwise-addition of taxa. We employed ML analyses using RAxML Black-Box webserver (http://www.ch.embnet.org/raxml-bb/, Stamatakis et al., 2008) with partioned model and bootstrapped with 100 replicates (BP_{ML}).

BI was performed using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). Two different criteria were used for the analyses: (1) separate nucleotide substitution models for each partition; and (2) a new method, named "Neutral Transitions Excluded" (NTE) for limiting the misleading effects of a reverse strand-bias in the data (Hassanin et al., 2005; Hassanin, 2006). In the NTE method, all neutral and quasi-neutral transitions were excluded from the original nucleotide matrix, as their nucleotide-substitution types are most likely to be affected by the bias. A GTR+I+G model (lset nst = 6) was used for first and second codon-positions, and a two-state substitution model (lset nst = 2) was used for third codon positions. In the case of all the Bayesian analyses, the Markov chain Monte Carlo (MCMC) were run for 10,000,000 generations, with tree sampling every 100 generations, to allow adequate time for convergence. Parameter convergence was achieved within two

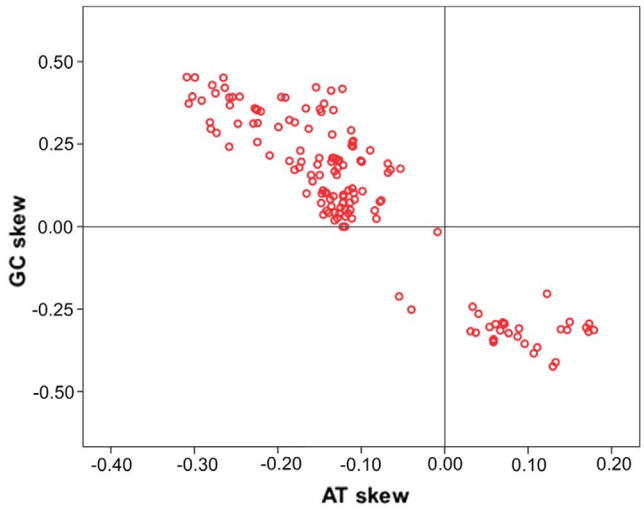

Fig. 1. Scatterplots of skews values calculated for whole plus strand of 140 molluscs mitochondrial genomes. All species fall into two groups: the first one includes eight species in Gastropoda and all species in Cephalopoda (typical patterns of skew group, TPS), presenting positive AT skews and negative GC skews; and the second one includes most species in Gastropoda and all species in Bivalvia (atypical patterns of skew group, APS), which are characterized by negative AT skews and positive GC skews.

million generations and the standard deviation of split frequencies was less than 0.01. All parameters were checked with Tracer v 1.5 (Drummond and Rambaut, 2007). After omitting the first 50,000 "burnin" tree, the remaining 50,000 sampled trees were used to estimate the 50% of majority rule consensus tree and the Bayesian posterior probabilities.

3. Result and discussion

3.1. Reversals of strand-compositional bias in molluscan mitochondrial genomes

AT and GC skew were determined for 140 molluscan complete genomes (plus strand) as a measure of the compositional asymmetry (Fig. 1, Supplementary Table 3). All species fall into two groups: the first one includes only thirty-four species (Typical patterns of skew group, TPS), i.e., eight species in Gastropoda and all species in Cephalopoda, presenting positive AT skews and negative GC skews; and the second one includes one hundred and ten species (Atypical patterns of skew group, APS), i.e., almost all species in Gastropoda and all species in Bivalvia, which are characterized by negative AT skews and positive GC skews, implying that these species have strand asymmetry reversal on the entire plus strand, with an excess of T relative to A and an excess of G relative to C. Despite the variations between species within both groups, there is a highly significant difference between these groups ($P < .0001$).

The customary explanation for strand asymmetry is that the deamination of A and C nucleotides occurs so much more frequently in single-stranded DNA than in double-stranded DNA (Rocha et al., 2006). Deamination of A nucleotide yields hypoxanthine, a base that pairs with C instead of T, while deamination of C nucleotide yields uracil, which can pair with A rather than G (Lindahl, 1993). Therefore, changes in strand-specific mutation rates may be associated with the location at which transcription begins in the mt genome, and with the length of time that a DNA strand remains single-stranded (Reyes et al., 1998).

An inversion of the control region contains the replication origin is expected to change the replication order of two mitochondrial DNA strands, resulting with time, in a complete reversal of strand compositional bias (Hassanin et al., 2005). These hypotheses have been demonstrated in two vertebrates (Fonseca et al., 2008), a echinoderms (Hassanin et al., 2005), a crustacean (Kilpert and Podsiadlowski, 2006) and ten insects (Wei et al., 2010) that the inversion of the control region explains the reversal of strand asymmetry. However, the gene order (and relative orientation) is variable and the control region is the most variable region in mitochondrial genome. It is not clear either whether the origin of replication is located in the major non-coding region (maybe the control region) of molluscs mt genomes, or whether this region is reversed in almost all species in Gastropoda and all species in Bivalvia.

3.2. Strand asymmetry on protein-coding genes

The AT and GC skews were calculated for all codon positions (P_{123}), third codon positions (P_3), and fourfold redundant third codon positions (P_{4FD}) for individual protein coding genes for each mitochondrial genome as a measure of the compositional asymmetry. At P_{123}, almost all genes coded on both the plus and minus strands have negative AT skew in APS group. This is the case in all species of TPS group (Supplementary Fig. 1A). Most mt genomes in APS group possess GT biased plus-strand genes, with a positive GC skew. However, the genes coded on minus strand showed the negative GC skew. The nucleotide composition of the TPS group displayed the opposite pattern of CG skew, with negative values in plus strand genes unlike other genes coded on the minus strand (Supplementary Fig. 1B and C). This result confirmed the assumption that two genes encoded by two opposite strands are expected to produce reverse strand compositional biases (Hassanin et al., 2005). The comparisons between AT and CG skews reveals that absolute values of CG skews are always significantly higher than that of AT skews ($P < .001$, paired sample *t*-test). This difference was also reported in six protein-coding genes of 49 metazoan mt genomes by comparing GC and AT skews at all codon positions (Hassanin et al., 2005), with the conclusion that CG skews are the best indicators of strand asymmetry.

At the P_3, especially the P_{4FD}, all genes in both APS and TPS group displayed similar patterns of AT and GC skews with the strand bias in their mt genome sequences. In other words, the taxa with reverse strand bias in their mt genome sequences also displayed atypical patterns of AT and GC skew at synonymous positions (Fig. 2). We tested the correlation between the sign of skew values on P_{4FD} and strand compositional bias of mt genome sequence using contingency table and chi-square test. The sign of both AT and GC skew values are associated with strand compositional bias of mt genome sequence, however, the chi-square values were lower for AT skew ($\chi^2 = 492.764$) than for GC skew (χ^2, 492.764 vs. 679.060, $P < .0001$). Here, skew values on individual genes for fourfold redundant third codon positions helped to explain the criterion that strand asymmetry is best reflected in the GC skew.

In all species, the absolute values of AT and GC skews on P_{4FD} are higher than those on P_{123} (Fig. 3, $P < .001$, paired sample *t*-test). Thus, the strand asymmetric base composition is stronger in weakly constrained sites. This is consistent with the hypothesis that the compositional asymmetry is particularly evident at synonymous codon positions of protein coding genes, which can freely alternate between all nucleotides without changing the resulting amino acid, are considered to have little or no affect on selection (Perna and Kocher, 1995; Reyes et al., 1998).

3.3. Effects of DNA asymmetric strand bias on amino acid composition

It has been shown for the nuclear genome that changes in the evolutionary patterns of amino acid substitution is the consequence of the corresponding changes in nucleotide content, e.g. GC content (Knight et al., 2001; Wang et al., 2004). However, although increasing mt genome sequencing efforts have led to the discovery of asymmetrical nucleotide biases among taxa, the relationship between asymmetrical nucleotide bias and potentially resultant amino acid bias has not been widely investigated within molluscs generally.

Given the contrasting patterns of mitochondrial gene strand

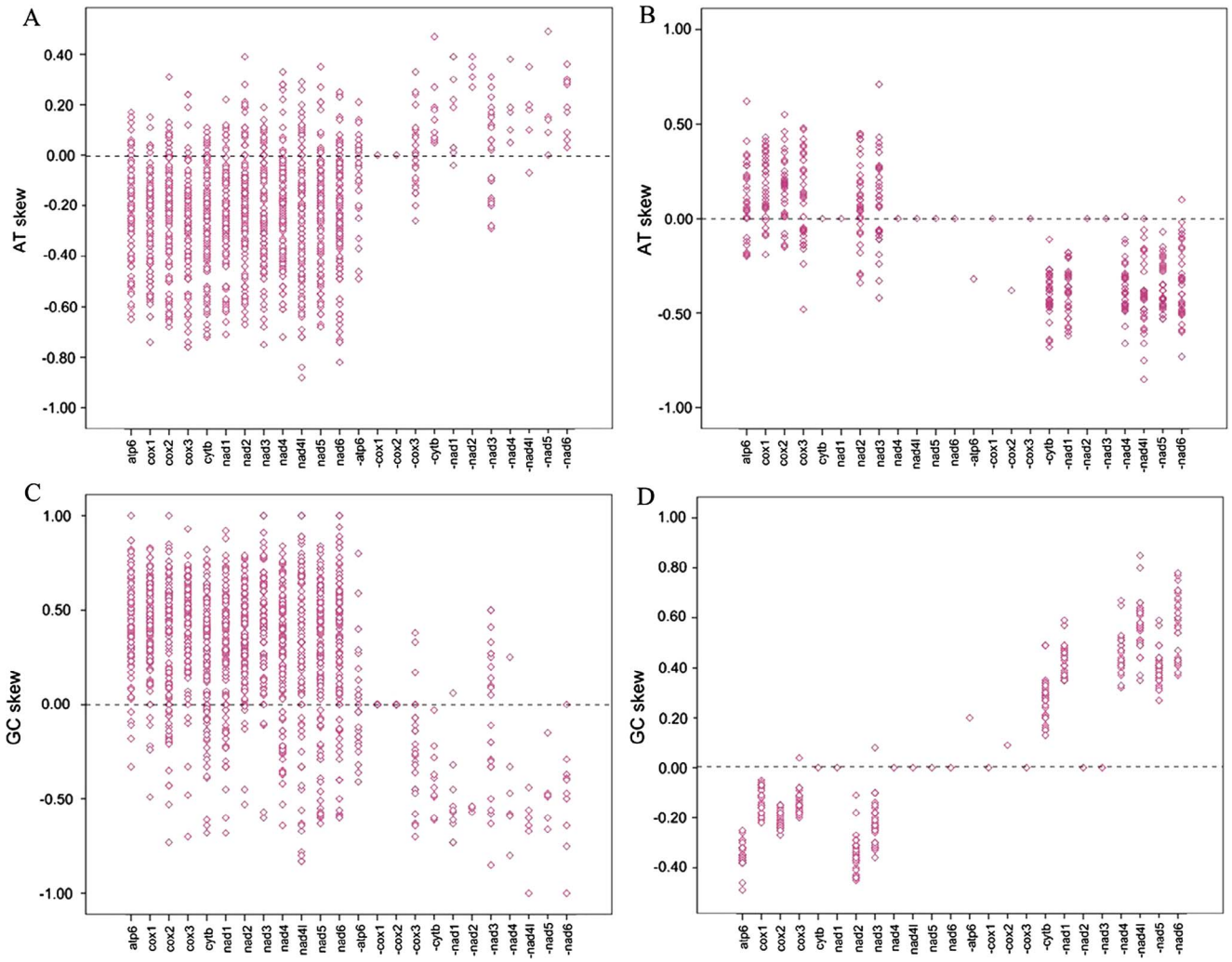

Fig. 2. AT and GC skews values of fourfold redundant third codon positions (P_{4FD}) in twelve individual protein-coding genes (excluding *atp8* gene) in molluscs mitochondrial genomes. (A). AT skews values calculated for P_{4FD} of individual protein-coding genes in atypical patterns of skew (APS) group. (B). AT skews values calculated for P_{4FD} of individual protein-coding genes in typical patterns of skew (TPS) group. (C). GC skews values calculated for P_{4FD} of individual protein-coding genes in atypical patterns of skew (APS) group. (D). GC skews values calculated for P_{4FD} of individual protein-coding genes in typical patterns of skew (TPS) group.

asymmetry in molluscs, we selected the same set of 6 conserved mitochondrial proteins (*atp6*, *cox1*, *cox2*, *cox3*, *nad2*, and *nad3*) coded on the plus-strand to investigate if there was a corresponding asymmetry between the two groups of molluscs in the frequencies of encoded amino acids. *Atp8* gene was excluded, because it was absent in most marine bivalve species, although several exceptions exist (Dreyer and Steiner, 2006; Wang et al., 2010; Wu et al., 2013; Gaitán-Espitia et al., 2016). The atypical patterns of skew (APS) group had a higher

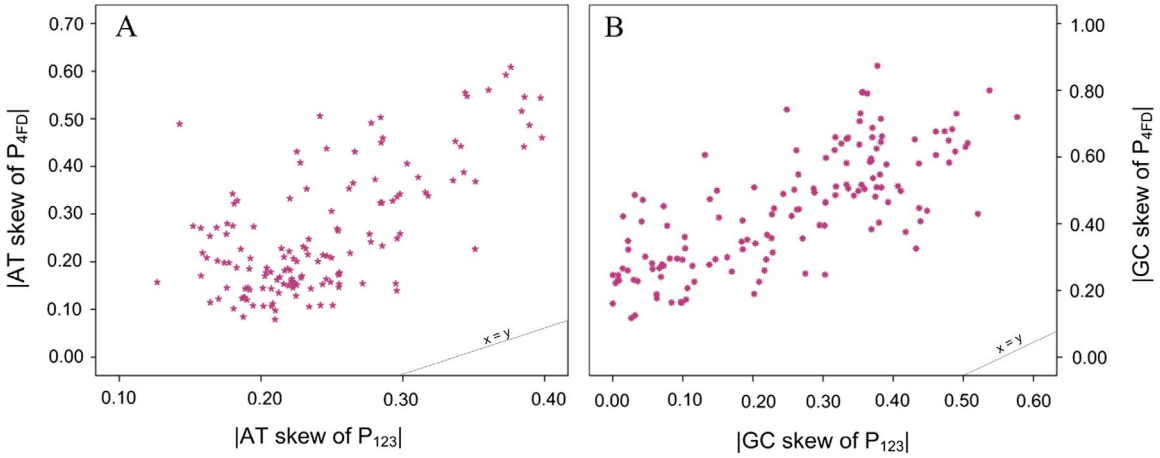

Fig. 3. Plot of the AT and GC skew on the third positions of fourfold degenerate codons (P_{4FD}) against the corresponding skew on all three codon positions (P_{123}) for each mtDNA analyzed. For both AT and GC skews, absolute values are presented.

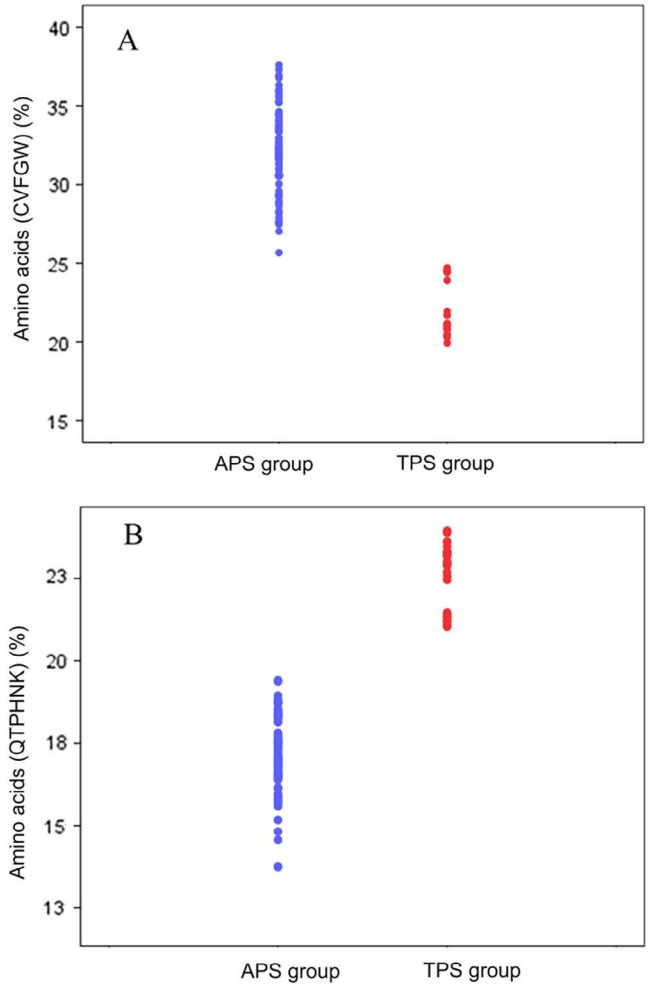

Fig. 4. The amino acids composition of mitochondrial protein-coding genes reflects the nucleotide strand symmetry. (A) The proportions of amino acids with GT-rich codons (Cysteine (C), Valine (V), Phenylalanine (F), Glycine (G), and Tryptophan (W)) are relatively high in atypical patterns of skew (APS) group (shown in blue) and relatively low in typical patterns of skew (TPS) group (shown in red). (B) The proportions of amino acids with CA-rich codons (Glutamine (Q), Threonine (T), Proline (P), Histidine (H), Asparagine (N), and Lysine (K)) are, in contrast, relatively low in atypical patterns of skew (APS) group (shown in blue) and relatively high in typical patterns of skew (TPS) group (shown in red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

abundance of amino acids encoded by GT-rich codons (Cysteine, Valine, Phenylalanine, Glycine, and Tryptophan), because of their positive GC skews and negative AT skews on the plus-strand (rich in G and T nucleotides) (Fig. 4A). On average, the difference of combined proportions of these five amino acids between the APS and TPS groups is approximately twofold. Conversely, amino acids with CA-rich codons (Glutamine, Threonine, Proline, Histidine, Asparagine, and Lysine) were less abundant in taxa with elevated GT richness than in typical taxa (Fig. 4B). Again, the average difference between the two groups of species is approximately twofold. Not only are the average differences in the predicted direction, but they are statistically highly significant ($P < .001$, *t*-test). Simultaneously, this significant differences ($P < .001$) also presented at the level of some individual amino acids, allowing to predict which codons and amino acids show clear responses to shifting nucleotide use. For example, APS group contains more than twice as many Valine and cysteine as do their orthologs in TPS group (Fig. 5A). On the other hand, the proportions of Asparagine and Threonine in APS group are approximately half the value observed in the TPS group orthologs (Fig. 5B). Thus, the differences in the patterns of strand asymmetry between the two strands of mtDNA significantly influenced the amino acid composition of the encoded proteins. These patterns have also been observed in the mitochondrial genomes of various metazoan groups in which mitochondrial strand asymmetry (measured as GC and AT skews) can have very large, predictable effects on the amino acid skew (Boore et al., 2004; Foster et al., 1997; Helfenbein et al., 2001; Min and Hickey, 2007).

Further tests were performed to verify the conclusion that the strand asymmetry at nucleotide level drive codon and amino acid usage rather than being a passive reflection of a preferred codon or amino acid usage. First, a larger strand asymmetry occurred at the synonymous codon sites (Supplementary Fig. 2), suggesting that the nucleotide skew was counterbalanced, to some extent, by functional constraint at the protein level. This is consistent with the fact that the codon usage may been altered with the changes at the third codon position, but the protein sequence remained the same. In other words, protein function was a constraint rather than a cause of the DNA strand bias (Min and Hickey, 2007). Secondly, in the TPS group, six gene (*Cytb*, *nad1*, *nad4*, *nad4l*, *nad5* and *nad6*) are encoded on the opposite strand from the other six genes and, as expected, the amino acids with CA-rich codon are more frequent than that of GT-rich. This pattern is similar to that of the proteins in the APS group rather than the other six proteins in TPS group. Both of these observations suggested that nucleotide asymmetry between the two strands of the mitochondrial genome is an artifact of selection (or mutation) and that this DNA bias causes a secondary effect at the level of protein composition.

3.4. Genome architecture and strand asymmetry

For the species characterized by a reverse strand asymmetry, it is necessary to find out how gene content or gene order changed. The mitochondrial genome architecture (AR) is defined as the order of the entire set of functional mt-encoded genes, which takes into account both gene content and gene order (Gissi et al., 2008). The variability in mt genome AR for every species was estimated by the "AR rate" (Table 1, Supplementary Table 4), according to the formulae $(N_{AR} - 1)/(N_{mtDNA} - 1) \times 100$. As shown in Table 1, AR rate values higher than 70, indicatives of a strong variability in genome AR, are observed in two order, e.g., Arcoida and Pterioida in class Bivalvia. Based on the analyses of the strand-compositional bias and the genome architecture in molluscs mt genomes, the taxa with a strong variability of genome AR are usually characterized by a reverse strand bias. However, species with reversal of strand asymmetry over the entire mitochondrial genome don't always have higher genome AR rates, e.g., species in Ostreoida and Unionoida in class Bivalvia, as well as Caenogastropoda and Neritimorpha in class Gastropoda. Although some of these groups are represented by a few sequences, the relatively conserved AR seems to be credible, as in some cases few AR differences were detected in the closely related species within that taxonomic group (Grande et al., 2002; Castro and Colgan, 2010). This result suggests a possible correlation between these two features and that same mechanism could be responsible for both the presence of the strand asymmetry and variability in gene number/order.

3.5. Phylogenetic relationships

3.5.1. Evidence for long-branch attraction artifacts

Although the analyses of mitogenomic sequence data usually lead to good resolution for phylogenetic relationships at low taxonomic levels, such as relationships between species, genera or even families. The usefulness of mtDNA sequences has been questioned for higher taxonomic levels such as relationships between orders, classes, or phyla (Curole and Kocher, 1999). The reversals of strand bias can be a crucial factor for explaining the difficulties encountered by many phylogeneticists for studying deep divergences with mtDNA sequences, as it can yielded longer branchs (more substitution) (Hassanin et al., 2005). What could be the consequences of such reversals of strand bias for

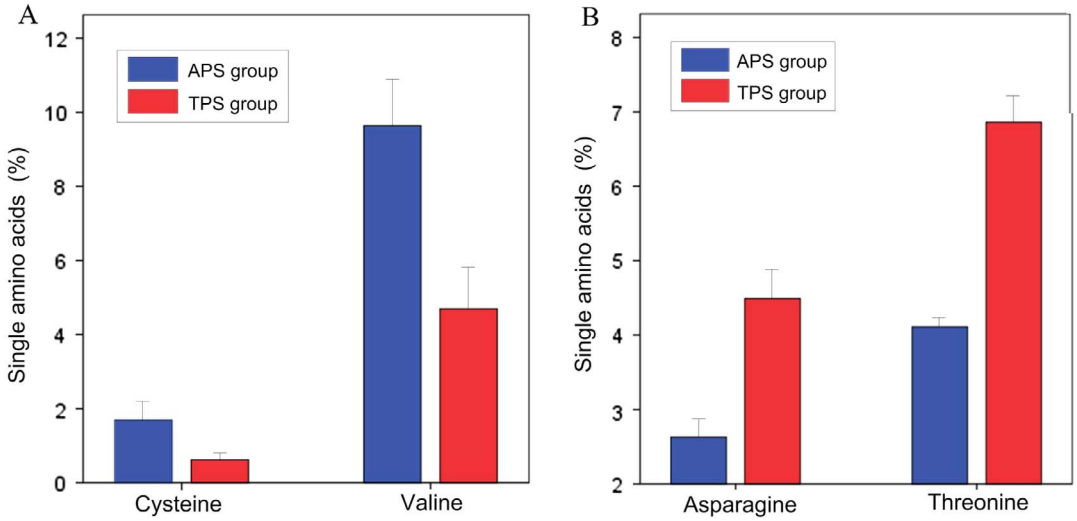

Fig. 5. The proportions of individual amino acids that were most affected by nucleotide strand asymmetry in atypical patterns of skew (APS) group and typical patterns of skew (TPS) group. (A) The proportions of Cysteine and Valine are high in atypical patterns of skew (APS) group proteins (shown in blue) and high in flatworm proteins (shown in red). (B) The proportions of Asparagine and Threonine are high in atypical patterns of skew (APS) group (shown in blue) and low in typical patterns of skew (TPS) group (shown in red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Distinct genome architecture and gene strand asymmetry of the analyzed molluscs mtDNAs.

NCBI classification		N_{mtDNA}	N_{AR}	AR Rate
Bivalvia				
	Arcoida	6	6	100.0
	Ostreoida	13	5	33.3
	Pectinoida	6	4	60.0
	Mytiloida	9	6	62.5
	Pterioida	2	2	100.0
	Veneroida	23	14	59.1
	Unionoida	5	2	25.0
Gastropoda				
	Heterobranchia	25	13	50.0
	Caenogastropoda	17	3	12.5
	Vetigastropoda	9	4	37.5
	Neritimorpha	4	2	33.3
Cephalopoda				
	Sepiida	10	1	0.0
	Teuthida	8	3	28.6
	Vampyromorpha	1	1	0.0
	Nautilida	2	1	0.0

Note: AR, genome architecture; the number and percentage of distinct genome architectures were calculated for the entire set of mt genes (protein-coding gene, tRNA, and rRNA). The variability of genome arrangement (AR rate) was calculated according to the formulae $(N_{AR} - 1)/(N_{mtDNA} - 1) \times 100$. The value is reported only for taxa with more than one complete genome.

phylogenetic inference?

Here, we conducted the ML and Bayesian analyses using nucleotides from all codon positions, which result in similar topologies of Mollusca (Fig. 6). The topology from parsimony analysis shows only minimal differences (Fig. 7). Within Bivalvia, the monophyly of subclasses Palaeoheterodonta (the group that includes freshwater pearl mussels), Heteroconchia and Pteriomorphia were evaluated, with Palaeoheterodonta grouped to all other autolamellibranchiates. This hypothesis is similar to the topology that has been recovered in previous mtDNA analyse (Gissi et al., 2008), but the relationships here are not at all congruent with current bivalve relationships based on sanger and transcriptome data (Bieler et al., 2014; Combosch et al., 2017; González et al., 2015). The internal resolution of Cephalopoda is in agreement with the current hypotheses (Allcock et al., 2011; Strugnell and Nishiguchi, 2007; Wilson et al., 2010). Cephalopods are again monophyletic with Nautiloidea sister to coleoids, comprising Decabrachia versus Octobrachia. Mitogenomic markers thus may be informative for resolving cephalopod relationships. Gastropods are the by far most diverse molluscan class, and also display greatest sequence heterogeneity (Stöger and Schrödl, 2013). In this study, the monophyly of gastropods cannot be tested, which split into two larger clusters, (1) Heterobranchia (with Nudipleura, lower Heterobranchia, Panpulmonata and Euopisthobranchia), (2) Caenogastropoda (represented by Neogastropoda, Littorinimorpha and Sorbeoconcha) together with Neritimorpha as sister to Vetigastropoda in a more derived position. The partial Gastropoda, namely Heterobranchia clustered with bivalves, while the second cluster, comprising Vetigastropoda, Neritimorpha and Caenogastropoda is sister group to Cephalopoda. These results were not consistent with other recent molluscan phylogenies, which evaluated and supported the monophyly of gastropod using both nuclear and transcriptome data (Kocot et al., 2011; Smith et al., 2011; Zapata et al., 2014). However, few non-gastropods molluscs were included (Zapata et al., 2014), and no Heterobranchia were included (Kocot et al., 2011; Smith et al., 2011) in these analysis.

When the asymmetric mutational constraints are reversed independently in several taxa, these taxa are expected to group together due to the long branch attraction (LBA) phenomenon (Felsenstein, 1978). This is what we observed in this study. Heterobranch gastropods form a well-supported but long branched clade that is pulled away from all other gastropods, but clustered with bivalves. This discrepancy may be due to an LBA artifact associated to the long branches exhibited by heterobranch and bivalve mt genomes (Grande et al., 2008; Stöger and Schrödl, 2013), suggesting that the phylogenetic position of Heterobranchia should be regarded with caution. Alternatively, the long branches rather reflect the accelerated evolutionary rates of the protein coding mitochondrial genes in heterobranch and bivalves. This is consistent with the hypothesis that Heterobranchia has high mitochondrial evolutionary rates that may result in a long-branch attraction artifact (Grande et al., 2008; Stöger and Schrödl, 2013; Arquez et al., 2014; Uribe et al., 2016).

3.5.2. Phylogenetic analysis under NTE model

In an attempt to overcome the confound phylogenetic inference caused by taxa with reverse strand bias, Hassanin et al. (2005) specified a Neutral Transitions Excluded (NTE) recoding scheme, which removes the effect of strand-bias by recoding bases at neutral and nearly-neutral positions as purines and pyrimidines (R/Y coding). This strategies have

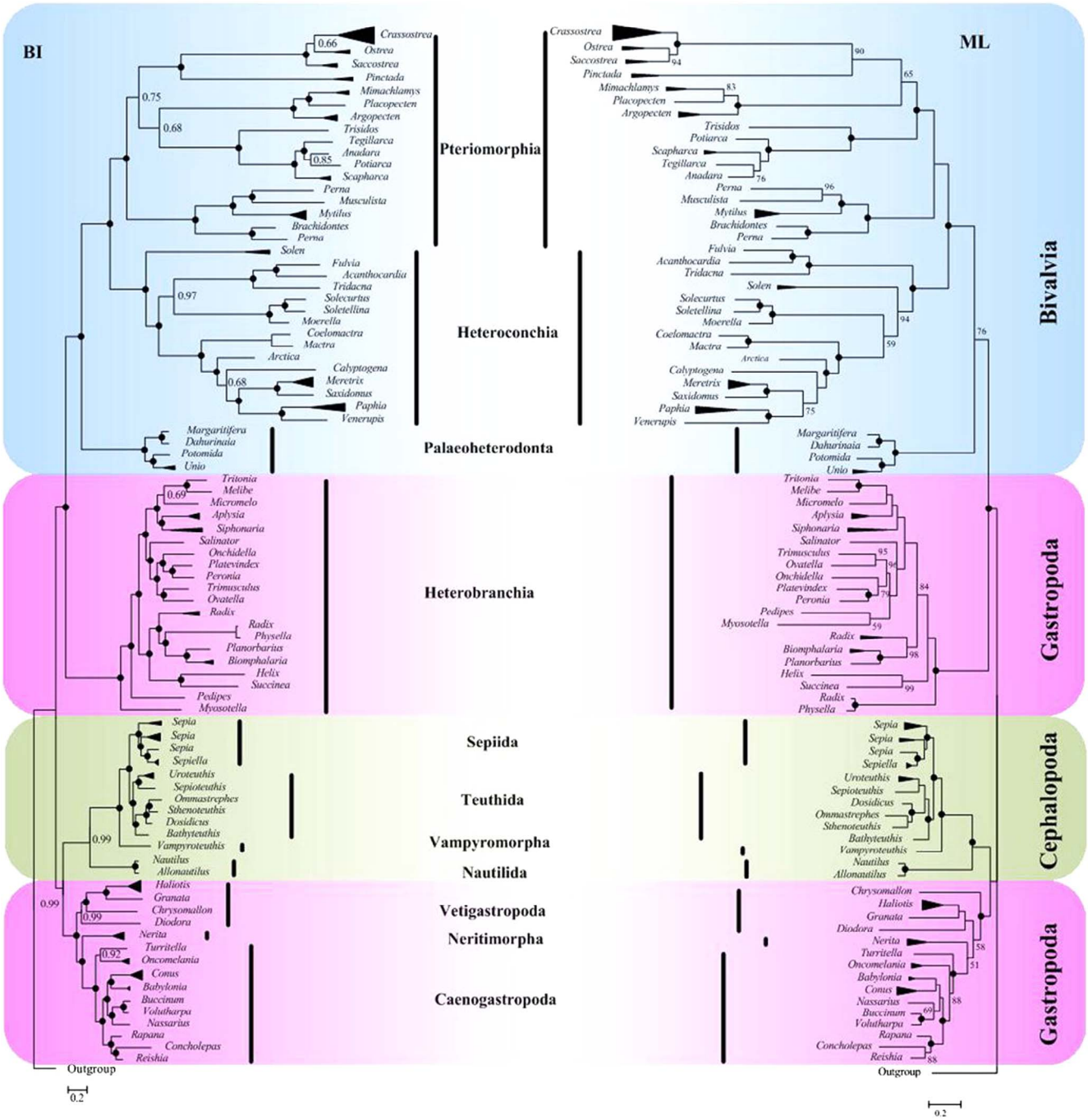

Fig. 6. Phylogenetic relationships of Mollusca based on nucleotides from all codon positions. The ML (A) and BI (B) phylograms are shown using GTR + I + G. Numbers at nodes are support values from ML (bootstrap proportions, BP) and BI (posterior probabilities, PP). Only BP > 50 and PP > 0.50 are listed. Filled circles represent nodes with BP = 100 and PP = 1.00. Colors indicate classic higher taxonomic ranks of Mollusca, with Bivalvia, Gastropoda and Cephalopoda marked in blue, pink and green, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

been implemented to help overcome the heterogeneous biases present in some arachnid mt genomes (Hassanin, 2006; Jones et al., 2007). However, in our analysis, an only slightly different topology were recovered both when the strand compositional bias was not accounted for and when the NTE method was used (Fig. 8). There is still no support for monophyly of gastropod molluscs.

It is difficult to know why this is, but we see two possible explanations. First, we think the processes such as long-branch attraction and heterotachy (site specific substitution rate changing through time) interacting with secondary structure may not be the only factors responsible for topology. Special attention for resolving deep nodes should be given to the relative arrangement of mitochondrial genes (e.g. Boore and Brown, 1995; Kurabayashi and Ueshima, 2000). Molluscs usually show accelerated rates of mitochondrial gene rearrangements and also known members of bivalves and some gastropod groups such as heterobranchs, which are aberrant and highly variable (Gissi et al., 2008; Simison and Boore, 2008; Stöger and Schrödl, 2013). Second, the NTE method (Hassanin et al., 2005) recommends RY recoding of the first codons positions of Leu (L1) and Phe (F), and of all three positions for Iso (I), Met (M), Ala (A), Thr (T), and Val (V). In present analysis, we found that the greatest differences in amino acid abundance among APS and TPS taxa are in Val and Cys, Asp and Thr.

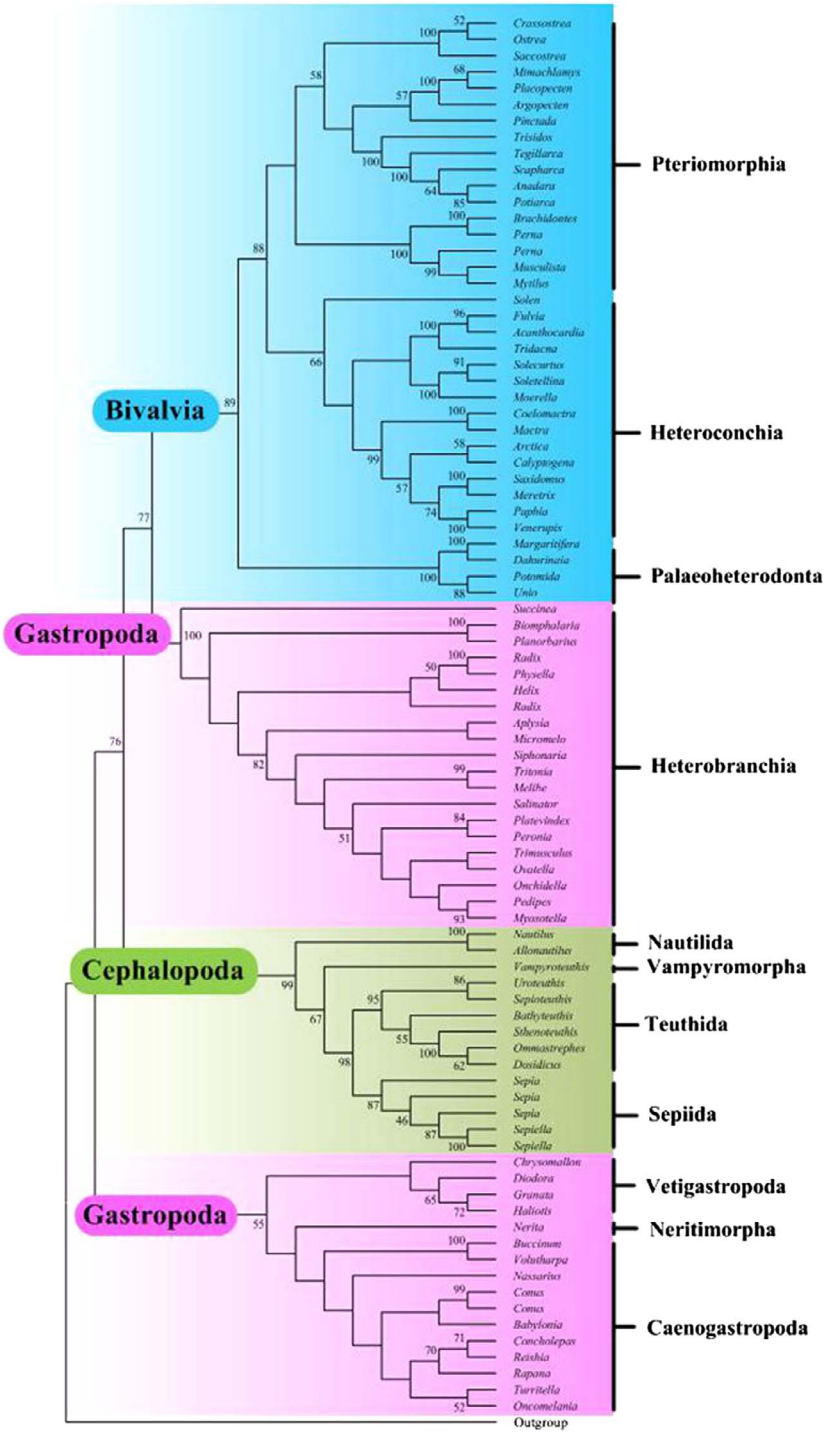

Fig. 7. Maximum parsimony topology of molluscs relationships. Bootstrap support values are indicated in each node. Colors indicate classic higher taxonomic ranks of Mollusca, with Bivalvia, Gastropoda and Cephalopoda marked in blue, pink and green, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, NTE recoding of Leu, Phe, Iso, Met and Ala may cannot account for this bias, making unnecessarily cause a loss of information useful for inferring relationships. An appropriate model of evolution for molluscs mtDNA should account for the actual changes among molluscs taxa, and the NTE method, although designed for metazoan mtDNA, may not be appropriate for molluscs mt genome data. We suggest that a thorough understanding of the molecular patterns and processes affecting the data in question is vital if more accurate phylogenetic models are to be explored.

Acknowledgments

This study was supported by research grants from National Natural Science Foundation of China (41276138), Fundamental Research Funds for the Central Universities, and Qingdao National Laboratory for Marine Science and Technology (2015ASKJ02).

Competing financial interests

The authors declare no competing financial interests.

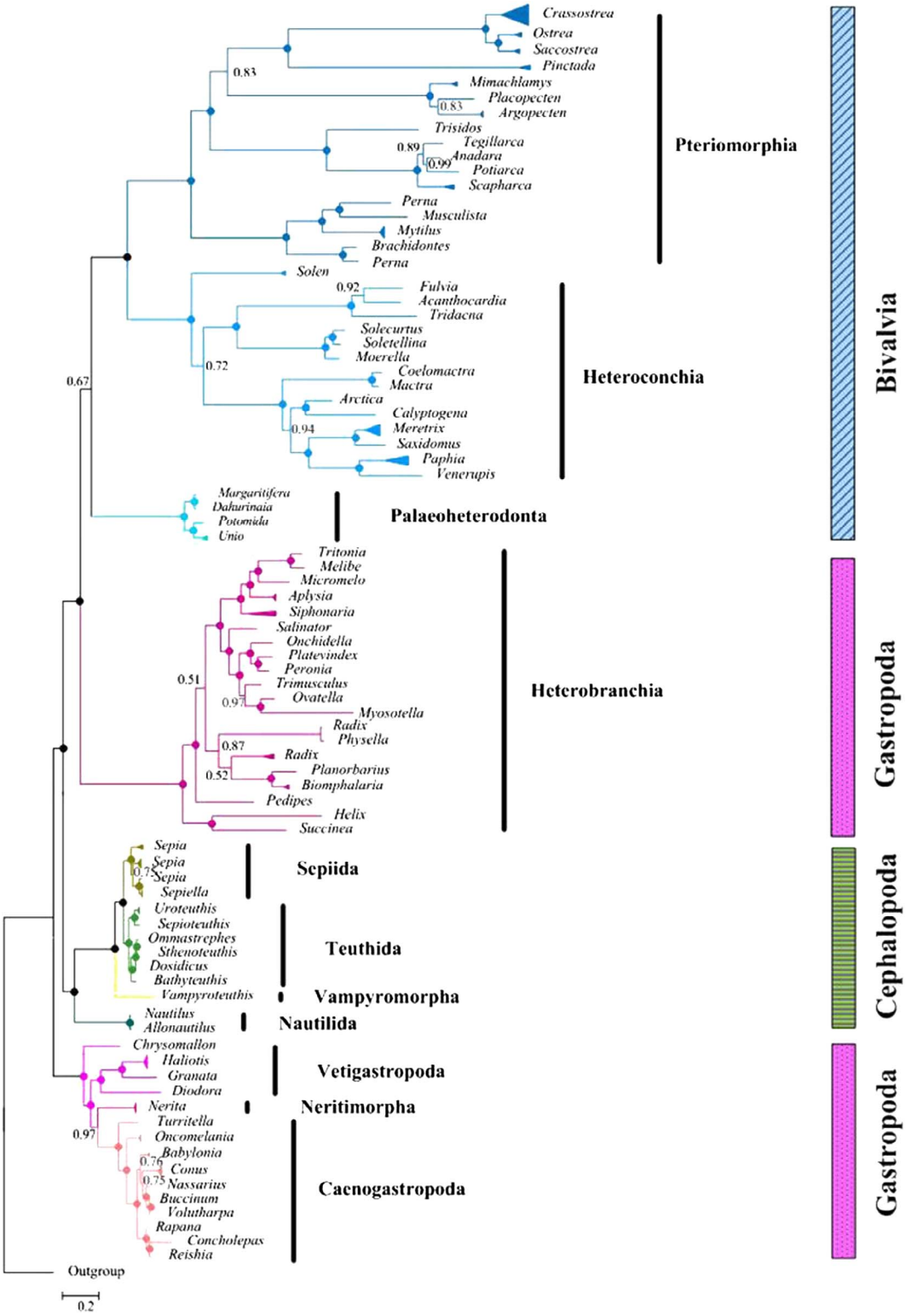

Fig. 8. Bayesian tree obtained by using the Neutral Transitions Excluded (NTE) recoded dataset. A GTR + I + G model (nst = 6; rates = invgamma) was applied to the first and second codon position and two substitution types (nst = 2) was applied to the third codon position. The values indicated on the branches correspond to posterior probabilities. Filled circles represent nodes with PP = 1.00.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.10.009.

References

Allcock, A.L., Cooke, I.R., Strugnell, J.M., 2011. What can the mitochondrial genome reveal about higher-level phylogeny of the molluscan class Cephalopoda? Zool. J. Linn. Soc.-Lond. 161, 573–586.

Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H.,

Staden, R., Young, I.G., 1981. Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.
Arquez, M., Colgan, D., Castro, L.R., 2014. Sequence and comparison of mitochondrial genomes in the genus *Nerita* (Gastropoda: Neritimorpha: Neritidae) and phylogenetic considerations among gastropods. Mar. Genom. 15, 45–54.
Avise, J.C., 2000. Phylogeography-The History and Formation of Species. Harvard University Press, Cambridge, Mass.
Ballard, J.W.O., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729–744.
Bernt, M., Bleidorn, C., Braband, A., Dambach, J., Donath, A., Fritzsch, G., Golombek, A., Hadrys, H., Jühling, F., Meusemann, K., Middendorf, M., Misof, B., Perseke, M., Podsiadlowski, L., von Reumont, B., Schierwater, B., Schlegel, M., Schrödl, M., Simon, S., Stadler, P.F., Stöger, L., Struck, T.H., 2013. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. Mol. Phylogenet. Evol. 69, 352–364.
Bieler, R., et al., 2014. Investigating the Bivalve Tree of Life-an exemplar-based approach combining molecular and novel morphological characters. Invertebr. Syst. 28, 32–115.
Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27, 1767–1780.
Boore, J.L., Brown, W.M., 1995. The complete DNA sequence of the mitochondrial genome of the annelid worm *Lumbricus terrestris*. Genetics 141, 305–319.
Boore, J.L., Medina, M., Rosenberg, L.A., 2004. Complete Sequences of the Highly Rearranged Molluscan Mitochondrial Genomes of the Scaphopod *Graptacme eborea* and the Bivalve *Mytilus edulis*. Mol. Biol. Evol. 21, 1492–1503.
Burger, G., Gray, M.W., Lang, B.F., 2003. Mitochondrial genomes: anything goes. Trends. Genet. 19, 709–716.
Cameron, S.L., Johnson, K.P., Whiting, M.F., 2007. The mitochondrial genome of the screamer louse *Bothriometopus* (Phthiraptera: Ischnocera): effects of extensive gene rearrangements on the evolution of the genome. J. Mol. Evol. 65, 589–604.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
Castro, L.R., Colgan, D.J., 2010. The phylogenetic position of Neritimorpha based on the mitochondrial genome of *Nerita melanotragus* (Mollusca: Gastropoda). Mol. Phylogenet. Evol. 57, 918–923.
Clayton, D.A., 1982. Replication of animal mitochondrial DNA. Cell 28, 693–705.
Combosch, D.J., et al., 2017. A family-level Tree of Life for bivalves based on a Sanger-sequencing approach. Mol. Phylogenet. Evol. 107, 191–208.
Curole, J.P., Kocher, T.D., 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Tree 14, 394–398.
Doucet-Beaupré, H., Breton, S., Chapman, E.G., Blier, P.U., Bogan, A.E., Stewart, D.T., Hoeh, 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.
Dreyer, H., Steiner, G., 2006. The complete sequences and gene organisation of the mitochondrial genomes of the heterodont bivalves *Acanthocardia tuberculata* and *Hiatella arctica*-and the first record for a putative Atpase subunit 8 gene in marine bivalves. Front. Zool. 3, 13.
Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Fonseca, M.M., Posada, D., Harris, D.J., 2008. Inverted replication of vertebrate mitochondria. Mol. Biol. Evol. 25, 805–808.
Foster, P.G., Jermiin, L.S., Hickey, D.A., 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44, 282–288.
Gaitán-Espitia, J.D., Quintero-Galvis, J.F., Mesas, A., D'Elía, G., 2016. Mitogenomics of southern hemisphere blue mussels (Bivalvia: Pteriomorphia): insights into the evolutionary characteristics of the *Mytilus edulis* complex. Sci. Reports 6, 26853.
Giribet, G., Okusu, A., Lindgren, A.R., Huff, S.W., Schrödl, M., Nishiguchi, M.K., 2006. Evidence for a clade composed of molluscs with serially repeated structures: monoplacophorans are related to chitons. Proc. Natl. Acad. Sci. U.S.A. 103, 7723–7728.
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.
González, V.L., et al., 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc. Roy. Soc. B Biol. Sci. 282, 20142332.
Grande, C., Templado, J., Cervera, J.L., Zardoya, R., 2002. The complete mitochondrial genome of the Nudibranch *Roboastra europaea* (Mollusca: Gastropoda) supports the monophyly of opisthobranchs. Mol. Biol. Evol. 19, 1672–1685.
Grande, C., Templado, J., Zardoya, R., 2008. Evolution of gastropod mitochondrial genome arrangements. BMC Evol. Biol. 8, 61.
Hassanin, A., Leger, N., Deutsch, J., 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst. Biol. 54, 277–298.
Hassanin, A., 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38, 100–116.
He, C.B., Wang, J., Gao, X.G., Song, W.T., Li, H.J., Li, Y.F., Liu, W.D., Su, H., 2011. The complete mitochondrial genome of the hard clam *Meretrix meretrix*. Mol. Boil. Rep. 38, 3401–3409.
Helfenbein, K.G., Brown, W.M., Boore, J.L., 2001. The complete mitochondrial genome of the articulate brachiopod *Terebratalia transversa*. Mol. Biol. Evol. 18, 1734–1744.
Jones, M., Gantenbein, B., Fet, V., Blaxter, M., 2007. The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions. Mol. Phylogenet. Evol. 43, 583–595.
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.
Kilpert, F., Podsiadlowski, L., 2006. The complete mitochondrial genome of the common sea slater, *Ligia oceanica* (Crustacea, Isopoda) bears a novel gene order and unusual control region features. BMC Genom. 7, 1.
Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2, 1.
Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., Schander, C., Moroz, L.L., Lieb, B., Halanych, K.M., 2011. Phylogenomics reveals deep molluscan relationships. Nature 477, 452–456.
Kurabayashi, A., Ueshima, R., 2000. Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. Mol. Biol. Evol. 17, 266–277.
Lavrov, D.V., 2007. Key transitions in animal evolution: a mitochondrial DNA perspective. Integ. Comp. Biol. 47, 734–743.
Lindahl, T., 1993. Instability and decay of the primary structure of DNA. Nature 362, 709–715.
Masta, S.E., Longhorn, S.J., Boore, J.L., 2009. Arachnid relationships based on mitochondrial genomes: Asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol. Phylogenet. Evol. 50, 117–128.
Min, X.J., Hickey, D.A., 2007. DNA asymmetric strand bias affects the amino acid composition of mitochondrial proteins. DNA Res. 14, 201–206.
Moreira, D., Philippe, H., 2010. Molecular phylogeny: pitfalls and progress. Int. Microbiol. 3, 9–16.
Morton, J.E., Yonge, C.M., 1964. Classification and structure of the Mollusca. Physiol. Mollusca 1, 1–58.
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. Mol. Biol. Evol. 25, 1253–1256.
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.
Rocha, E.P.C., Touchon, M., Feil, E.J., 2006. Similar compositional biases are caused by very different mutational effects. Genome Res. 16, 1537.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Scouras, A., Smith, M.J., 2006. The complete mitochondrial genomes of the sea lily *Gymnocrinus richeri* and the feather star *Phanogenia gracilis*: signature nucleotide bias and unique nad4L gene rearrangement within crinoids. Mol. Phylogenet. Evol. 39, 323–334.
Simison, W.B., Boore, J.L., 2008. Molluscan evolutionary genomics. In: Ponder, W.F., Lindberg, D.R. (Eds.), Phylogeny and Evolution of the Mollusca. University of California Press, London, pp. 447–461.
Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C., Rouse, G.W., Giribet, G., Dunn, C.W., 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480, 364–367.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758–771.
Stöger, I., Schrödl, M., 2013. Mitogenomics does not resolve deep molluscan relationships (yet?). Mol. Phylogenet. Evol. 69, 376–392.
Strugnell, J., Nishiguchi, M.K., 2007. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) inferred from three mitochondrial and six nuclear loci: a comparison of alignment, implied alignment and analysis methods. J. Mollus. Stud. 73, 399–410.
Swofford, D.L., 2003. PAUP∗. Phylogenetic analysis using parsimony (∗and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Tanaka, M., Ozawa, T., 1994. Strand asymetry in human mitochondrial DNA mutations. Genomics 22, 327–335.
Uribe, J.E., Kano, Y., Templado, J., Zardoya, R., 2016. Mitogenomics of Vetigastropoda: insights into the evolution of pallial symmetry. Zool. Scr. 45, 145–159.
Wang, H.C., Singer, G.A., Hickey, D.A., 2004. Mutational bias affects protein evolution in flowering plants. Mol. Biol. Evol. 21, 90–96.
Wang, X., Wang, J., He, S., Mayden, R.L., 2007. The complete mitochondrial genome of the Chinese hook snout carp *Opsariichthys bidens* (Actinopterygii: Cyprinifonnes) and an altemative pattem of mitogenomic evolution in vertebrate. Gene 399, 11–19.
Wang, H., Zhang, S., Xiao, G., Liu, B., 2010. Complete mtDNA of the *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 5, 256–264.
Wei, S.J., Shi, M., Chen, X.X., Sharkey, M.J., van Achterberg, C., Ye, G.Y., He, J.H., 2010. New views on strand asymmetry in insect mitochondrial genomes. PLoS ONE 5, e12708.
White, T.R., Conrad, M.M., Tseng, R., Balayan, S., Golding, R., de Frias Martins, A.M., Dayrat, B.A., 2011. Ten new complete mitochondrial genomes of pulmonates (Mollusca: Gastropoda) and their impact on phylogenetic relationships. BMC Evol. Biol. 11, 1.
Wilson, N.G., Rouse, G.W., Giribet, G., 2010. Assessing the molluscan hypothesis Serialia (Monoplacophora+ Polyplacophora) using novel molecular data. Mol. Phylogenet. Evol. 54, 187–193.
Wu, X., Li, X., Yu, Z., 2013. The mitochondrial genome of the scallop *Mimachlamys senatoria* (Bivalvia, Pectinidae). Mitochondrial DNA 26, 242–244.
Xu, X., Wu, X., Yu, Z., 2012. Comparative studies of the complete mitochondrial genomes of four *Paphia* clams and reconsideration of subgenus *Neotapes* (Bivalvia: Veneridae). Gene 494, 17–23.
Zapata, F., Wilson, N.G., Howison, M., Andrade, S.C.S., Jörger, K.M., Schrödl, K.M., Schrödl, M., Goetz, F.E., Giribet, G., Dunn, C.W., 2014. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proc. Roy. Soc. B. 281, 20141739.