[2, 3]. After that, mitochondria coevolved with di erent hosts and underwent both neutral modifications and adaptive responses that led to the diversity observed today in mitogenomes [4]. In bilaterians, mitogenomes were considered to be extremely compact and normally organized into a single circular molecule ranging in size from 14 to 20 kb [5]. Bilaterian mitogenomes typically contain the same set of 37 genes (13 protein-coding genes encoding di erent subunits of enzyme complexes for the oxidative phosphorylation (OXPHOS) system, 2 ribosomal RNAs (rrnS and rrnL), 22 transfer RNA genes) and no introns [5–7]. In the last few years, high-throughput sequencing techniques and extensive sampling for phylogenetic and population genetic studies have accelerated the sequencing of mitogenomes and uncovered the great diversity of structural features [8]. An increasing number of mitogenomes seem to deviate dramatically from typical bilaterian mitogenomes and present wide variation in genome size, and many of them are much larger than 20 kb. Many molluscs, especially bivalves, display an unusual amount of variation in mitogenome structure and size, even among closely related species [6, 9, 10].

Arcidae, known as Ark shell or blood cockles, are an economically-important group of bivalves and have a long evolutionary history, dating back to the Lower Ordovician ~450 Mya [11]. Interestingly, mitogenomes of Arcidae species both within and between species reveal a high variability in size, ranging from 19 to 56 kb in length [12]. For example, reported mitogenomes of are 46,985 bp [13] and 48,161 bp [14], the recently published mitogenome is 45,697 bp [15], which are 2–3 times larger than other bilaterians. e largest Arcidae mitogenome comes from (46.7–56.2 kb) and is the largest bilaterian mitogenome yet recorded, out of approximately 86,900 mt-DNAs from more than 11,600 species [8, 12]. In addition, large mitogenomes are also found in sea scallop (31-41 kb) [16] and the clavagellid mussel (32 kb) [17].

e large genome sizes in ark shells and sea scallop are not primarily a result of duplications of control region sequences and coding sequences like model organisms, but rather the expansion of unassigned regions (i.e., non-coding regions that are functionally unassigned) [16, 18]. A previous analysis [19] of 2656 complete mitogenomes showed that some bivalves have a proportion of unassigned regions (URs) that are significantly di erent from all other groups and show the highest median percentage of URs in Metazoans. According to the mutation pressure theory [20], fast evolving organelle genomes experience more selection pressure for genome reduction, but some bivalve mitogenomes seem to contradict this theory. In some Arcidae species, URs account for more than 50% of the entire mitogenomes. Tandem repeats, inverted repeats and transposable elements in unassigned regions (URs) have been shown to contribute to the large size of these mitogenomes [12, 13, 16], but it does not mean they are the main cause of huge expansion in URs. Data from previous studies show that repeat families and transposable elements are not the main components of large URs, which only account for 6–31% of URs in different Arcidae species, though they have a significantly positive correlation with mitogenome size [12, 13]. is suggests that there are other components influencing the size of URs. One possible explanation is that retention of the URs in bivalve mitogenomes is caused by the presence of functional sequences and/or structures. However, to date, much of the work on URs of Arcidae has focused on nucleotide-level analyses to observe sequence characteristics (e.g., tandem repeats, inverted repeats), mitogenome expansion remains poorly understood and needs further study with di erent perspectives.

Mitochondria are long known for bioenergetics, but they also have novel non-OXPHOS-related adaptations and functions [21]. With the increasing number of published mitogenomes, non-standard gene contents have been found in di erent animal groups, and additional mitochondrial protein-coding genes have been identified and annotated in mitogenomes of metazoan, particularly in invertebrates. For example, additional mitochondrial protein-coding genes were first discovered in the octocoral [22], which was a homolog of and hypothesized to originate either through bacterium or viral infection by horizontal gene transfer [23, 24]. In cnidarians, sponges and placozoans, protein-coding genes with non-OXPHOS functions (e.g., ,) have been also reported [25, 26]. Surprisingly, nine additional mtDNA-encoded protein genes have been described in humans [27–30], one of which is a 75 bp ORF in the mitochondrial 16S rRNA that acts as a neuroprotector, an antiapoptotic agent, and a cytoprotector [31, 32]. ese discoveries indicate that there are additional functional sequences in mitochondria, maybe related to its diverse functions.

Moreover, multiple ORF sequences have also been found in the mitogenomes of bivalves. A novel ORF was discovered with no sequence- or domain-based homology to mitochondrial genes in the mitogenome of pearl-lip oyster but has domain-based homology to the nuclear genome [33]. Mitochondrial ORFans (open reading frames having no detectable homology and no known function) also have been identified in marine and freshwater bivalves (Mytiloida, Nuculanoida, Unionoida, and Veneroida) with doubly uniparental inheritance (DUI) of mitochondrial DNA [34, 35]. In these cases, products are exported from the

organelle and may be involved in functions other than energy production [34–40]. ese studies indicate that traditional bivalve mitochondrial non-coding regions have sequences or unassigned regions that potentially perform biological functions. e structure of some special ORFs in the mitogenome of (Arcidae) have been briefly investigated [41], but the origin of Arcidae ORFs in large URs remain unclear. In addition, a fundamental question regarding the size of mitogenomes in Arcidae bivalves is whether there are ORFs in large URs that perform functions and have a connection with expanded size of mitogenomes.

Here, we sequenced and annotated five new mitogenomes (four and one) and assembled a mitogenome of from NCBI data (SRX8857271). Six transcriptomes of were sequenced and analyzed to identify conserved functional ORFs. Multiple samples from the same species were used to detect intraspecific variation in mitogenome length and the presence of ORFs. To better understand how URs expand and evolved in Arcidae, we present a comparative analysis of 32 complete mitogenomes (6 new assemblies and 26 published assemblies from NCBI) of Arcidae species to highlight both unique features and characteristics shared among di erent species, with an emphasis on characterizing large URs and ORFs. en, we investigated the origin and duplication of ORFs and their correlation with mitogenome expansion, and particularly with the expansion and function of mitochondrial large URs in Arcidae.

Results

Mitogenome assembly, annotation and features

Complete mitogenomes sequences of four , one and one had sequence lengths more than 40 kb (Table 1). Four new mitogenomes sequences varied in size from 44,327 bp to 48,560 bp, close to previously published mitogenome (46,985 bp) [13]. e length of mitogenome reported here was 54,157 bp, slightly smaller than the previously reported (56,170 bp) [12], and is 46,362 bp long. and mitogenomes vary dramatically in length within species. All mitogenomes consisted of 12 protein-coding genes (all taxa lacked 8), two ribosomal RNA genes (and) and 27–33 tRNA genes (Additional file 1: Table S1). has never been found in Arcidae species [12, 18, 41, 42]. All protein-coding genes and rRNA genes in the six mitogenomes are encoded on the same strand and share the same gene order. In addition, a duplication of was observed in all six mitogenomes (Additional file 1: Table S2), located between and 2. e two copies have di erent length: the 2 is 666–720 bp long (221-239aa), while the 2-b is 1179–1431 bp long (392-476aa). is indicates that 2-b have acquired an extension after 2 duplication. Four di erent start codons (ATG, ATA, ATT, GTG) were observed but most protein-coding genes start with the codon ATG, and stop with the TAA and TAG codons. e organization of tRNAs was variable across the six mitochondrial genomes sequenced here. All mitogenomes are composed of four major segments: two coding regions and two major unassigned regions (Fig. 1). ere is little variation in length of coding regions and great variation in URs (Fig. 2). ese newly sequenced complete mitogenomes were deposited in GenBank (Accession numbers: OM807131-OM807136).

Characterization of unassigned regions in mitogenomes

All available mitogenomes of , , and were characterized by large unassigned regions separated into two principal blocks. e first block (UR1) was located between 2 and 6, and the second block (UR2) was located between 2 and 1. We refer to these as “shared large unassigned regions”. An assessment of shared URs in nine mitogenomes (Table 2) revealed that the length of UR1 is relatively stable between 9831 and 10,120 bp except in where it was 14,034 bp. UR2 length was highly variable from 15,167 to 28,331 bp, and the intergenic DNA (URs between genes in coding blocks) ranged from 1507 to 2925 bp. Overall, the total length of URs was 28,844 to 41,101 bp, accounting for 65.1–73.2% of these nine mitogenomes. Tandem repeats in URs showed significant variation in number (2–26 copies) and sequence length (1-273 nt) (Additional file 1: Table S3). e total length of tandem repeats varied from 759 bp to 4942 bp and takes up only 2–15% of URs. In addition, an examination of all Arcidae complete mitogenomes showed that unassigned regions (i.e., repeats, ORFs) were highly variable and responsible for expansion and variation in Arcidae mitogenomes (Additional file 1: Table S4). In comparison, a higher proportion of unassigned sequences was observed in the ark shell of (>60%) and (>50%), both exceeding 16 kbp, and thus there was a strong positive correlation between mitogenome size and proportion of unassigned region.

Novel ORFs in mitochondrial unassigned regions

TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki) predicted eleven ORFs in mitochondrial unassigned regions that might code proteins from the transcriptome. e analysis of mtDNA transcriptome expression (Fig. 3) showed that 12 mitochondrial coding genes have higher transcription level than all ORFs but ORF8 and ORF21

Table 1 Mitochondrial (mt) genomes analyzed in this study, including newly assembled mitogenomes and those from Genbank

Species	Subfamily	Length (bp)	SRA	Locality
New mt genomes				
Scapharca broughtonii (1)	Anadarinae	44,333		Qingdao, Shandong, China
Scapharca broughtonii (2)	Anadarinae	44,327		Qingdao, Shandong, China
Scapharca broughtonii (3)	Anadarinae	46,191		Qingdao, Shandong, China
Scapharca broughtonii (4)	Anadarinae	48,560		Qingdao, Shandong, China
Scapharca cornea	Anadarinae	46,362		Philippines
Scapharca kagoshimensis (1)	Anadarinae	54,157		Qingdao, Shandong, China
Species	Family/Subfamily	Length (bp)	GenBank acc. no.	Publication
GenBank mt genomes				
Anadara crebricostata	Anadarinae	36,671	MN316632	Kong et al., [12]
Anadara transversa	Anadarinae	18,780	MN326817	Kong et al., [12]
Anadara vellicata	Anadarinae	34,147	KP954700	Sun et al., [42]
Lunarca ovalis	Anadarinae	19,620	MN366010	Kong et al., [12]
Potiarca pilula	Anadarinae	28,386	KU975162	Sun et al., [43]
Scapharca broughtonii (5)	Anadarinae	48,161	KF667521	Hou et al., [14]
Scapharca broughtonii (6)	Anadarinae	46,985	AB729113	Liu et al., [13]
Scapharca globosa	Anadarinae	33,405	MN366011	Kong et al., [12]
Scapharca gubernaculum	Anadarinae	45,697	MN061840	Sun et al., [15]
Scapharca inaequivalvis	Anadarinae	45,859	MN366012	Kong et al., [12]
Scapharca kagoshimensis (2)	Anadarinae	56,170	MN366013	Kong et al., [12]
Scapharca kagoshimensis (3)	Anadarinae	46,713	KF750628	Sun et al., [18]
Tegillarca sp.	Anadarinae	50,104	MN366016	Kong et al., [12]
Tegillarca granosa	Anadarinae	31,589	KJ607173	Sun et al., [41]
Tegillarca nodifera	Anadarinae	38,672	MN366014	Kong et al., [12]
Arca navicularis	Arcinae	18,004	MN326818	Kong et al., [12]
Arca zebra	Arcinae	44,651	MN366003	Kong et al., [12]
Barbatia lima	Arcinae	17,479	MN366005	Kong et al., [12]
Barbatia virescens	Arcinae	24,871	MN366006	Kong et al., [12]
Trisidos semitorta (1)	Arcinae	19,461	MN366015	Kong et al., [12]
Trisidos semitorta (2)	Arcinae	19,613	KU975161	Sun et al., [43]
Cucullaea labiate (1)	Cucullaeididae	25,845	KP091889	Feng et al., [44]
Cucullaea labiate (2)	Cucullaeididae	20,481	MN366007	Kong et al., [12]
Glycymeris formosana	Glycymeridídae	19,027	MN366008	Kong et al., [12]
Glycymeris yessoensis	Glycymeridídae	17,903	MN366009	Kong et al., [12]
Arcopsis adamsi	Noetiidae	18,716	MN366004	Kong et al., [12]
Outgroup				
Mizuhopecten yessoensis	Pectinidae	20,964	FJ595959.1	Wu et al., [45]
Pinctada maxima	Pteriidae	16,994	NC_018752.1	Wu et al., [46]
Crassostrea gigas	Ostreidae	18,224	AF177226.1	N/A

have a similar transcription level to . . had the highest transcription level of all PCGs and ORFs. In comparison, ORF104 and ORF127 showed a very low transcription level, which was considerably lower than other ORFs. e mapped read counts and TPM values of PCGs and ORFs have been recorded in Additional file 2. e ORFs shared the same location and order in four mitogenomes (Additional file 1: Table S1). All were located in URs on the heavy strand (as all standard coding genes) of mitogenomes, including six in UR1 and five in UR2. e eleven ORFs showed remarkable conservation in all samples: their start and stop codons were the same, respectively, except ORF127 in (1), ORF21 in (2) and ORF8 in (3) (Table 3). e longest length of ORFs was 1983 bp (ORF87, 660aa), the shortest length of ORFs was 238 bp (ORF21, 79aa). Notably, the ORF8 was variable

in its length (561 bp to 636 bp). Although has not been previously reported in Arcidae, the lengths of ORF21 (238 bp) and ORF103 (244 bp) are close to that of found in other bivalves (102–339 bp) and these ORFs may be candidates for . Moreover, we found potential ORF duplication events in four mitogenomes. Amino acid sequence comparisons using a combination of sequence and position similarity revealed that there are five additional ORFs that share a large degree of similarity within the eleven ORFs (Additional file 1: Table S5). Results showed that 47% amino acid identities were observed between ORF8 and ORF11, 39% for ORF8 and ORF7, 30% for ORF8 and ORF78. e highest identity (67% amino acid identity) was observed between ORF87 and ORF127. ORF104 had at least 28% amino acid identities with other three ORFs. ese findings showed that ORFs have high similarity, suggesting that duplication events have occurred between them. However, no significant amino acid sequence similarity was detected with known proteins for the 14 new lineage-specific ORFs using BLAST [47] and PSI-BLAST against NRDB and UniProt (Additional file 1: Table S6).

To establish whether these ORFs were taxonomically restricted to or if they were an evolutionary feature of ark shell, we screened for the presence of 16 ORFs (11 predicted ORFs and 5 duplicated ORFs, see Table 3 and Additional file 1: Table S5) in 32 mitogenomes of Arcidae (Table 1) with BLAST (Additional file 1: Table S7 and Additional file 3). None of the ORFs were similar to mitochondrial PCGs. Complete ORFs and their duplications could only be annotated in 13 mitogenomes, including 6 and 1 species (Additional file 1: Table S8 and Fig. 2), suggesting that most of the ORFs were specific to the lineage (except for). With one exception, ORF21 was annotated in 12 species, which had a wider distribution. All 16 ORFs were verified in the mitogenomes of and except for (3). Although the mitogenome of (3) have a long URs, only five complete ORFs could be annotated, which may be due to poor assembly quality of the mitogenome. e result showed that 15 ORFs were found in the mitogenome of except for ORF 11. In the mitogenomes of and , 14 ORFs were found. Only ORF8, ORF21 and ORF86 were found in mitogenome. Most of ORFs were unique to the genus and , but the partial sequence (about 120aa) of ORF87 was similar to fragments of other genera mitogenomes (Additional file 1: Table S7), which suggests that ORF87 may have had a wider distribution in Arcidae (Fig. 2). Sequence comparisons both within and between lineage-specific ORFs revealed high variability in length, but some ORF lengths were conservative such as ORF10 and ORF11 (Additional file 1: Table S8). e total length of all ORFs identified in was accounting for 23–26% of mitogenomes and 35–40% of URs, indicating ORFs are one of the main components of the large URs. In addition, many large ORFs (> 2000 bp) were found in Arcidae mitogenomes (Additional file 1: Table S9).

Ka/Ks analysis for putative novel Arcidae mitochondrial ORF proteins

To estimate the degree of selection (either neutral, positive, or purifying) and genetic conservation on 12 protein coding genes and 16 ORFs, the number of nonsynonymous substitutions per nonsynonymous sites (Ka) relative to the number of synonymous substitutions per synonymous sites (Ks) was calculated. According to our results (Fig. 4 and Additional file 1: Table S10), the mean value for Ka was also di erent between protein-coding genes and ORFs (0.0137 vs. 0.0624; < 0.001), suggesting that the ORFs accumulated more non-synonymous mutations. Twelve protein-coding genes and all ORFs were found to be under strong purifying selection (Ka/Ks < 1). e Ka/Ks of protein-coding genes was between 0.0470 ( ) to 0.1447 (4L). A low Ka/Ks has been a common finding for mtDNA-encoded protein genes in animals and is explained by the elimination of mildly deleterious mutations. e ORFs had a wide range of Ka/Ks ranging from 0.1267 (ORF103) to 0.615042 (ORF40), but the Ka/Ks of some ORFs (ORF5, ORF103, ORF127) were close to some protein-coding genes (4L, 5). e mean value of Ka/Ks in ORFs (0.2834) was significantly higher than that of protein-coding genes (0.0874) ($P < 0.001$), suggesting that ORFs have been under less selective constraints than mitochondrial protein-coding genes. In addition, the level of ORF sequence

(See figure on next page.)

Fig. 1 Maps of the six mitogenomes sequenced in this study. The corresponding species name and length are given inside each genome map. The outer ring comprises all standard and putative coding sequences, identified with the following color code: blue, genes encoding electron transport chain and ATP-synthase subunits; yellow, tRNA genes; green, rRNA genes (see Additional file 1: Table S1 for details). All the genes are encoded on the same strand. The middle ring represents GC content (dark grey). The inner ring represents scale

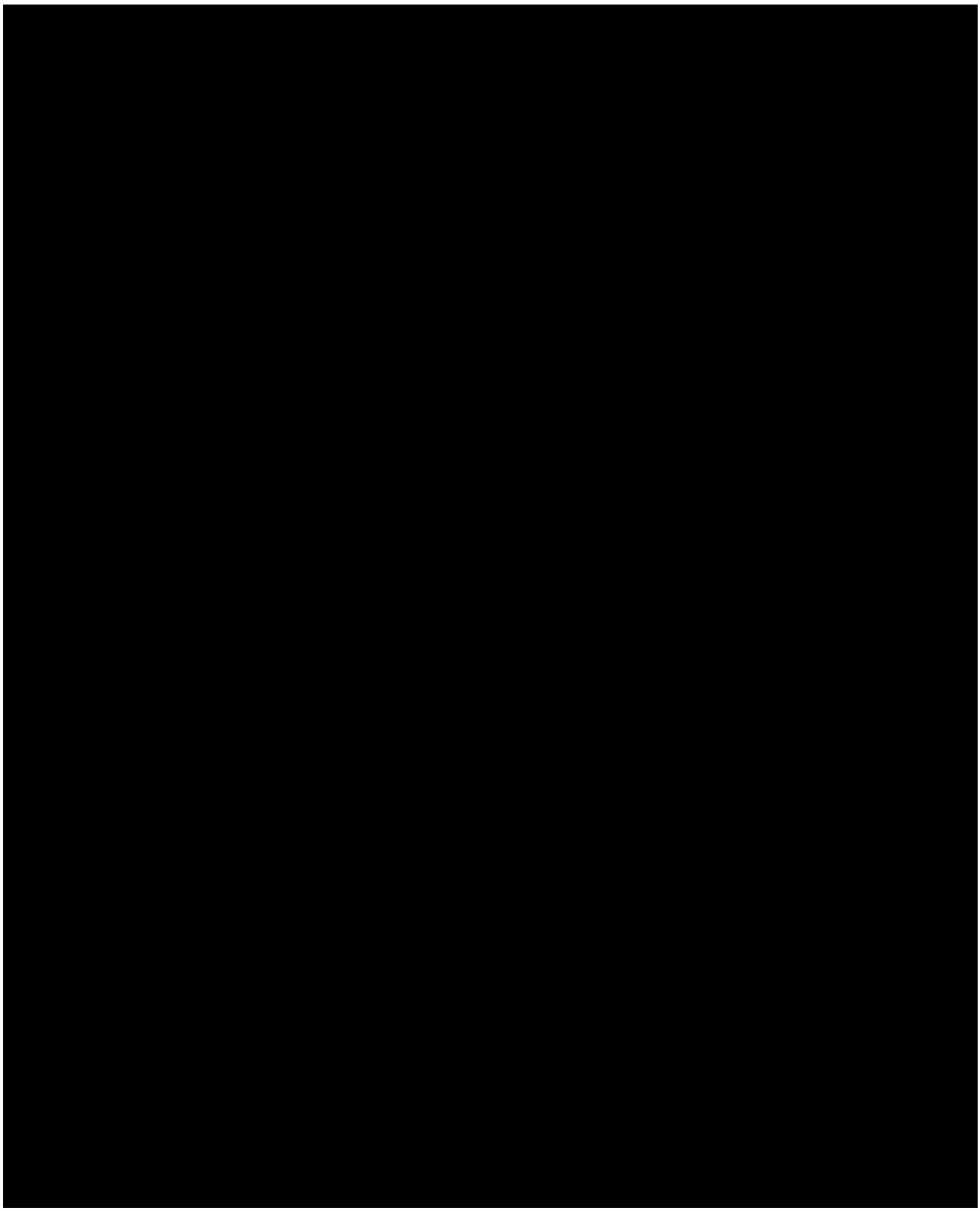


Table 2 Detailed information of unassigned regions of Arcidae mitogenomes. The unit of length is the bp

Species	Mitogenome size	The total length of URs	Proportion of URs	UR1 location	UR1 size	Proportion of UR1	UR2 location	UR2 size	Proportion of UR2
Scapharca broughtonii (1)	44,333	28,908	65.21%	18,057–27,994	9938	22.42%	1–861436,903–44,333	16,045	36.19%
Scapharca broughtonii (2)	44,327	28,844	65.07%	17,914–27,851	9938	22.42%	1–847136,760–44,327	16,039	36.18%
Scapharca broughtonii (3)	46,191	30,806	66.69%	18,033–27,968	9936	21.51%	1–859036,753–46,191	18,029	39.03%
Scapharca broughtonii (4)	48,560	32,984	67.92%	25,845–35,677	9833	20.25%	1–16,53244,063–48,560	21,030	43.31%
Scapharca broughtonii (5)	48,161	33,004	68.53%	9314–19,144	9831	20.41%	27,798–48,161	20,364	42.28%
Scapharca broughtonii (6)	46,985	31,809	67.70%	9314–19,146	9833	20.93%	27,673–46,985	19,313	41.11%
Scapharca kagoshimensis (1)	54,157	38,394	70.89%	38,928–48,778	9851	25.66%	2805–29,551	26,747	69.67%
Scapharca kagoshimensis (2)	56,170	41,101	73.17%	30,713–40,833	10,120	18.02%	1–21,32949,169–56,170	28,331	50.44%
Scapharca kagoshimensis (3)	46,713	32,982	70.61%	9395–19,345	9951	21.30%	28,023–46,713	18,691	40.01%
Scapharca cornea	46,362	30,708	66.24%	22,735–36,777	14,034	30.27%	1–13,57144,767–46,362	15,167	32.71%
Scapharca gubernaculum	45,697	30,079	65.82%	9164–21,106	11,943	26.14%	29,095–45,697	16,603	36.33%
Scapharca globosa	33,405	18,631	55.77%	14,724–17,921	3198	9.57%	1–437226,555–33,405	11,223	33.60%
Anadara crebricostata	36,671	21,991	59.97%	23,496–27,840	4345	11.85%	1–593336,253–36,671	6352	17.32%
Anadara vellicata	34,147	20,659	60.50%	9130–9424	295	0.86%	18,084–34,147	16,064	47.04%

predict the function of ark shell mitochondrial ORFans. Results obtained for all ORFans with all programs for protein function prediction were summarized in Additional file 1: Table S13, which included the most frequent categories of hits for molecular functions, biological processes and cellular components for the mitochondrial ORFans [51, 52], and detailed motifs and domains information (HHpred, [53]; I-TASSER). Overall, the most common hits for all ORFs were proteins involved in oxidoreductase activity, nucleic acid or protein binding (e.g., helicase/hydrolase activity) and metal ion binding (e.g., nickel cation/cobalt ion binding). Some other hits were proteins with membrane association and transporter activity, for example involved in transport across membrane, establishment of protein localization, and most of all involved in intracellular transport (e.g., ORF10, ORF7, ORF104, ORF106, ORF21). Some ORF proteins pointed to a role in ATP binding, for example in cellular macromolecule metabolic process (e.g., ORF127).

In particular, most sequences analyzed returned predictions that the proteins were involved in oxidoreductase activity and metabolic process, and the predicted subcellular localizations for these ORFs were di erent, with some being membranes and organelles (endoplasmic reticulum and nucleus) and some being soluble outside the cell (Additional file 1: Table S14). For ORF104, the highest probability matches included proteins that have a role in obsolete oxidoreductase activity, acting on

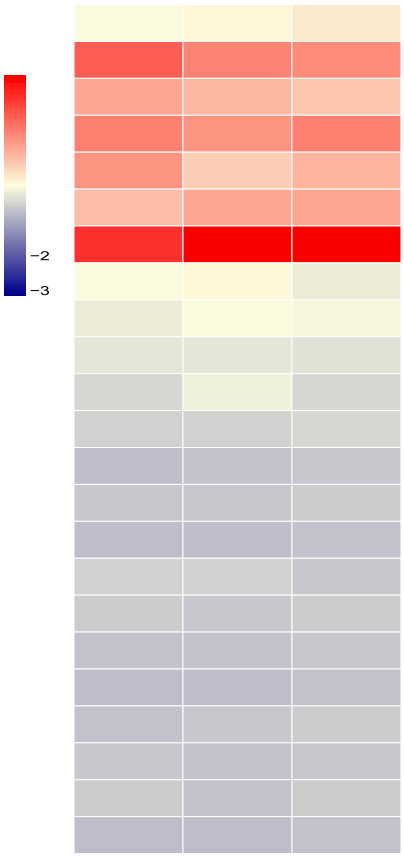
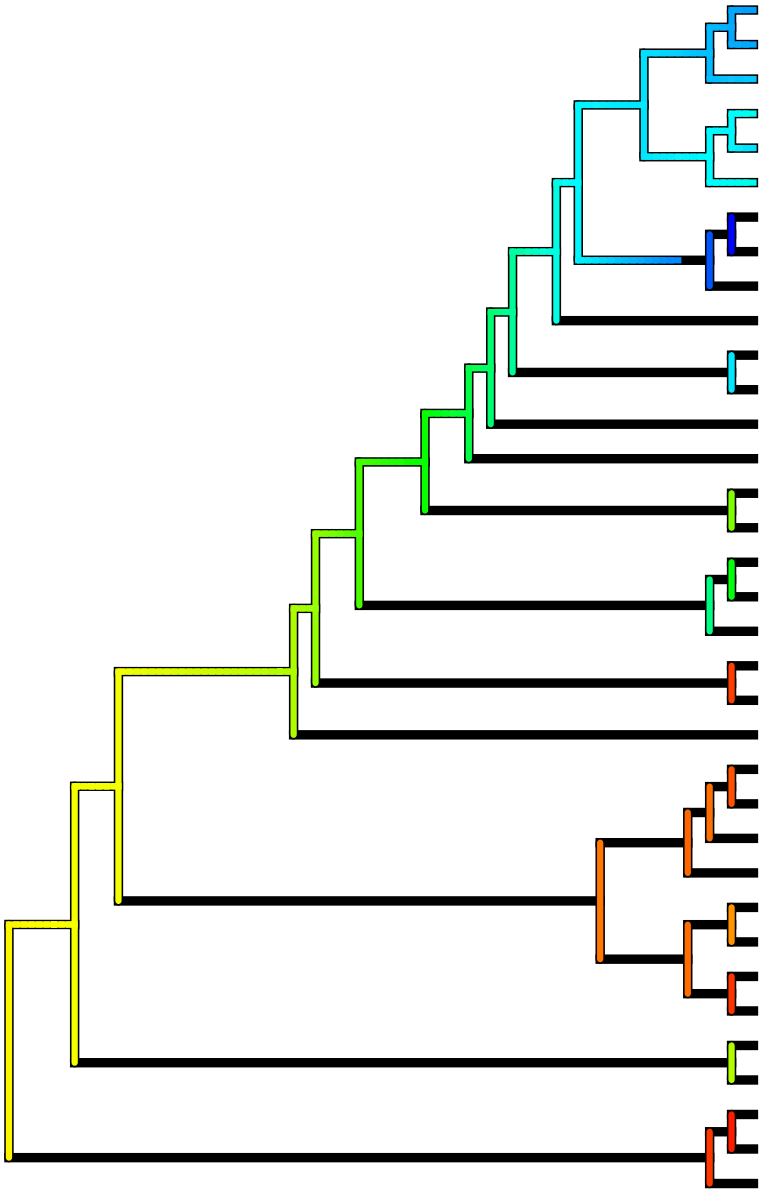


Fig.Rates yof

third lineage. , the only representative of , was found to nest within the polyphyletic Arcidae as the sister taxon of the clade. Within Anadarinae, and were found to be polyphyletic. ese results are consistent with previous studies [12, 44, 54–56]. Arcoidea, which includes , , , and , , formed a clade that was well-supported in the ML analysis (bootstrap support value = 100%), whereas the previous ML analysis [12] for Arcoidea was not well-supported (bootstrap support value = 56%). Ancestral state reconstruction indicated that the evolution of the mitogenome size has undergone di erent changes across di erent arcoid lineages (Fig. 6). A medium-sized mitogenome (23.8 kb) was estimated to be the ancestral state of Arcoidea. Mitogenome expansion was apparent in Anadarinae species, whereas genome contraction has occurred in , , , and . In addition, the clade, which encompasses species, , , , and , also have a slight contraction. Notably, multiple expansion takes place in genus , but an independent expansion in .

Discussion

In this study, we present multiple lines of evidence supporting the functionality of ORF at amino acid and nucleotide levels. ese results support novel features for Arcidae mitochondrial genomes: the presence of additional, lineage-specific, mtDNA-encoded proteins with potential functional significance. We discuss the possible origin of ORFs including mitochondrial, nuclear and viral origins. In addition, the insertion and duplication events of these ORFs play an important role in multiple expansions of Arcidae mitogenome, which may provide significant insights on how bilaterian mitochondrial genomes evolve in terms of size variation, gene complement, and gene organization.



Novel ORFs in Arcidae mitogenomes and their possible origins

Mitogenomes of Arcidae species with large URs have novel lineage-specific ORFs which do not show significant amino acid sequence similarity to known proteins in the database, but they have special secondary structures and meaningful hits for function prediction. e ORF amino acid sequences are conserved among di erent mitogenomes and found in extra-genic regions, always inside the large URs. Traditionally, URs are generally regarded as vestiges of pseudogenes generated by random deletions after gene duplication [34, 57–59], while our further examination of the two shared large URs in Arcidae reveals two categories of sequences. e first

category contains sequences exhibiting many features typically associated with mitochondrial control regions, such as presence of repeat units and sequences that can form hairpin structures and stem–loop. e second category contains sequences that possess many ORFs of considerable length. Previous research [12] on the first category found significantly positive correlations between Arcidae mitochondrial genome size and proportion of tandem repeats and proportion of inverted repeats. However, the ORFs identified turned out to be one of the main components of the large URs rather than repeat families and transposable elements in our study. Some ORFs were conserved in the lineage containing Scapharca and (Fig. 2), suggesting that they may have emerged before the speciation of lineage and . Mitochondrial URs of some ark shells are probably inserted in multiple independent evolutionary events [12], and if so these lineage-specific ORFs may arise from independent insertion events.

Previous studies [35, 38–40, 60] suggested mitochondrial ORFs could originate from di erent processes: (1) the duplication and subsequent modification of extant mitochondrial genes, (2) transfer of DNA from nuclear genomes to mitochondrion, (3) the insertion of viral sequences into the host mitogenome. In our study, the ORF sequences analyzed do not show homology with any known Arcidae mitochondrial protein, so they unlikely originated from recent duplication events from existing mtDNA genes. e ORF sequences we identified in were not found in a high-quality nuclear genome of [61] (Additional file 5), indicating they may not have been transferred from nuclear genomes. Moreover, in our analysis, hits similar to ORF sequences are mainly from proteins of bacteria, fungi, parasites and viruses (Additional file 1: Table S6 and Table S13). Since most of hits have no significant similarity or probability, they are not su cient to determine the origin of the ORFs. e results showed that 9 of 16 ORFs in are possibly involved in the process of viral entry into host cell and viral release from host cell, which provide clues to speculate on a possible viral origin. However, BLAST results including all Arcidae species showed that most of the ORF hits are to non-viral sequences, and the probability of some viral hits (<60%) were low. Hits with high probability values (E-values <0.005) are bacterial or metazoan proteins, suggesting that other organisms or other processes [62] may be the source of these ORF genes. Consequently, our results do not support a viral origin of Arcidae ORFs, although previous studies [38, 40] have suggested that some bivalve mitochondrial ORFs may originate from viruses.

In Arcidae mitogenomes, if ORFs have experienced rapid evolution since their origin, they may diverge to the extent that homology to mitochondrial proteins, nuclear sequences and viral sequences is not discernable. e ka/ks results showed that some ORF sequences contained mostly non-synonymous mutations (Additional file 1: Table S10), indicating the rapid evolution of these mitochondrial ORF genes. Fast rate of evolution may erase evidence of ORF sequence similarities (homology) among species, so we cannot fully exclude the possibility that the ORFs are derived from mitochondrial duplications, nuclear genome or even viral sequences. Of these possible origins, mitochondrial duplication interpretation is more plausible. Gene duplication is thought to be the most common mechanism underlying the origin of most novel genes [63]. In our study, large-scale duplication events (both mitochondrial genes and ORFs) were found, suggesting that duplications are prone to occur in Arcidae mitogenome. Also, this explains why many identified ORFs have TM-helices and the most common hits of them involved in metabolic process. In considering possible alternatives, these ORFs may originate from duplication of mitochondrial genes, but also call for further studies to investigate their origin.

The insertion and duplication of mitochondrial novel ORFs and implications for UR size evolution

According to ancestral state reconstruction (Fig. 6), the mitogenome size of the common ancestor of Arcidae is relatively small (i.e., <20 kb) like in most metazoans. is result suggests Arcidae mitogenomes have experienced multiple expansions (insertion and duplication events) with some lineages forming very large mitogenomes. For example, UR1 (regions between 2 and 6) is only 295 bp in the mitogenome of (Table 2). However, we found that the UR1 of mitogenome is 3197 bp, which contains the ORF sequences similar to ORF86 and ORF8. Based on phylogenetic relationships and mitogenome structure of Arcidae (Fig. 5), the inserted sequence may be the origin of the UR1 of lineage and . Subsequently, the expansion of UR1 is obvious in other species (reaching above 10 kbp) (Table 2) because of the duplications of some ORFs (Fig. 5). In bivalve, many mitochondrial gene duplication events have been found such as 2 duplication in the oyster genus (Bivalvia, Ostreidae) [46] and 2 duplication in several DUI bivalve species [64, 65]. In our results, mitogenomes contained a duplication of the 2 gene, named 2-b (Fig. 5 and Additional file 1: Table S2), a feature that has been also observed in the mitogenome of other closely related species (lineage and , except for). Coincidentally, ORF duplications were found in the species that have 2 duplication with exception of .

For example, according to ORF distribution and phylogenetic relationship (Fig. 5 and Additional file 6), ORF8, ORF7, ORF11 and ORF78 (>30% identity at the protein level) might originate from the same ORF gene and the ORF7 gene in could have been the original copy. In addition, duplication of ORF5, ORF21, ORF104, ORF46, ORF86 and ORF87 were also found in di erent Arcidae mitogenomes, which suggest that a large-scale duplication event has occurred rather than just 2 duplication. Meanwhile, these duplicated ORF sequences make mitogenomes larger. Most of the ORFs from duplication were over 500 bp in length, which adds at least 8217 bp to the mitogenomes of . For example, ORF87 and its duplication of ORF127 are both over 1800 bp, which have a huge impact on mitogenome size. In conclusion, the duplication of 2 and ORFs cause Arcidae mitogenomes to expand further. We think that duplication events of ORFs play an important role in multiple subsequent expansions of Arcidae mitogenome.

However, most of ORFs we identified in are only found in and (Fig. 2). Possible explanation is that the expansion of and was the result of independent insertion and duplication events, which results in the ORF sequences di erent from other genera. To further explore expansions caused by ORFs, we reinvestigated all Arcidae mitogenomes and found 54 large ORFs (LORFs) of unknown structure and function in 18 Arcidae mitogenomes (Additional file 1: Table S9). ese LORFs range in length from 933 to 5187 bp, and the average length is 2200 bp. In the mitogenomes of , all LORFs were found in UR2, indicating that the emergence of LORFs may be related to UR2 expansion. Interestingly, many LORFs were also found in and mitogenomes, and some of them have a fragment of amino acid sequences similar to the regions of ORF87 in (Fig. 2 and Additional file 1: Table S7), which implies that these LORFs might have a common origin with ORF87 in an earlier expansion event. Because of large lengths, LORFs were easily disrupted by mutations and harder to maintain. erefore, we think that the existence of LORFs might be significative and can provide clues to explore the independent expansions of other Arcidae mitogenomes in future.

Predicted functions for ark shell mitochondrial ORFs

In our study, there are multiple lines of evidence indicating potential functionality of 14 novel lineage-specific ORFs in , likely as expressed proteins. In other taxa, lineage-specific ORF genes are involved in important adaptive processes and key biological functions [36, 38]. Herein, multiple lines of evidence suggest these UR ORFs are functional in nature. Transcriptome analysis indicated that the nine ORFs (Table 3 and Fig. 3) are transcribed in mitogenomes of . Secondly, the ka/ks analysis showed that all the ORFs are under purifying selection (Fig. 4 and Additional file 1: Table S10). irdly, because novel ORFs do not show significant similarity to known proteins, we performed multiple analyses of their structure to predict the function. Our results for secondary structure prediction show that most of the ORFs are predicted to have functional domains (Additional file 1: Table S11), which is a significant support for identifying these ORFs as protein-coding genes. For example, we observe that four conserved TM-helices are present in ORF40 proteins of . In addition, ORF40 returned hits to proteins with membrane association (e.g., proteins involved in tail-anchored membrane protein insertion into ER membrane), and the predicted subcellular localizations with DeepLoc for ORF40 are also endoplasmic reticulum and membrane. e ORF127 protein is very long in length and has two stable TM-helices among di erent Arcidae species. Our prediction results, together with transcriptome and ka/ks analysis, showed that these lineage-specific ORF proteins that occur in Arcidae mitogenomes may have underlying functions. For example, ORF21 and ORF103 may be candidates for , which has not been annotated in many bivalves [6] and is not reported in Arcidae. e two ORFs are approximately the same length as and are transcribed in . ey are under strong purifying selection and their relative solvent accessibility is similar to the gene (Additional file 7). ORF21 is conserved at the amino acid level and has a broader distribution than ORF103 (Fig. 2). ORF103 is predicted to have TM-helices, but ORF21 does not. Compared to ORF21, we believe that ORF103 is a more likely candidate for because the TM-helix is important for function. However, ORF21 and ORF103 do not show homology with any bivalve atp8 genes. erefore, we cannot determine if either of these two ORFs is an gene homolog.

Moreover, based on the functional prediction analysis, we speculated that some of novel ORF proteins in mitochondrial URs of Arcidae may have acquired new functions. A previous study showed that ORFans in bivalves with DUI may have a viral origin and be involved in the maintenance of sperm mitochondria during embryo development [38, 40]. Our results for molecular function prediction show that the ORFs have di erent functions hits, the most common hits involved in metabolic process (e.g., ATP association activity, oxidoreductase activity, glucan 1,4-alpha-glucosidase activity). Several Arcidae species are limited locomotive and more tolerant to hypoxia, such as S. kagoshimensis, which lead to low metabolic rate [43, 66, 67]. e large mitogenome size

in these bivalves may be correlated with their metabolic rates because the relaxed selective constraints of mitogenomes may be energy-related [12, 68]. In addition, genes participating in ATP and lipid metabolism under selection were found to be important in thermal adaptation for oysters [69]. Lineage-specific genes could participate more in lineage-specific adaptation [70], therefore, we speculated that the functions of lineage-specific ORF genes are related to low metabolic rate or thermal adaptation, which may provide new insights into the function of large URs in Arcidae mitogenomes.

Finally, the possibility that ORFs may be pseudogenes and do not perform any function cannot be ruled out. According to the results, some ORFs are unstable in the number of TM-helices such as ORF104, ORF106 and ORF46. e number of TM-helices in a given ORF vary among Arcidae species, indicating that these ORFs are not conserved at the secondary structure level. e results suggest a possibility that ORFs may have di erent adaptations in various Arcidae species. Another possibility is that the ORFs could be pseudogenes or in the process of pseudogenization after insertion and duplication events. In many cases duplicated genes are subject to pseudogenization [1], which appears to be the most likely fate for mitochondrial gene duplications. But compared to pseudogenes, the ORFs we identified are more conserved at the nucleotide sequence level and some of them may be transcriptionally active, so they are likely to be in the process of pseudogenization but more research is needed.

Dynamic changes in intraspeci□c mitogenome size

Di erent mitogenomes in the same Arcidae species vary dramatically in length. e previous studies have demonstrated that tandem repeats are potentially a main factor leading to variation of intraspecific mitogenome size [12, 16]. For example, mitogenomes of ranges in size from about 47 kb to ~ 50 kb due to variation in the number of tandem repeats [13]. e four mitogenomes we assembled also have many tandem repeats (Additional file 1: Table S3). However, variation of these tandem repeat size does not fully explain the variation (44,327–48,560 bp) of the URs of mitogenomes because the latter is larger. By observing URs length (Table 2), UR1 in and are almost the same length, and UR2 is highly variable and responsible for the variation of mitogenomes in length. We propose that variation of the length of and mitogenomes may be caused by incomplete assembly of the UR2 because of their complex content. Large-scale repeat sequences are di cult to sequence using conventional Sanger and short-read sequencing methods [71]. eoretically, repeats extend beyond read length, mitogenome assemblies are limited within the boundaries of repetitive elements [72]. Although long-range PCR can be used to amplify DNA regions of several kilobases, sequencing through repetitive regions often results in ambiguous and/or erroneous sequence reads as a consequence of self-priming of randomly-amplified repeat-segments, chimeras and/or jumping PCR artefacts [73, 74]. For example, the mitogenome (KF750628) were assembled using long-PCR into a circle [18], but some protein-coding genes and ORFs are fragmented, which suggest a low-quality assembly. Hence, future studies can focus on long-read sequencing for Arcidae mitogenome assembly, which can achieve read lengths of 80 kb to > 1 Mb, enabling repetitive and structurally complex DNA elements to be resolved with confidence [75, 76].

Conclusions

In this study, we found 14 special ORFs in the large unassigned regions in mitogenomes of . Interestingly, these putative additional proteins have also been found in other species of genus . We present multiple lines of evidence supporting the functionality of ORFs at amino acid and nucleotide levels and discuss their possible origin. ese results support novel features for Arcidae mitochondrial genomes: the presence of lineage-specific mitochondrial ORFs with transcriptional activity and potential functional significance. Moreover, our study reveals that the insertion and duplication events of ORFs play an important role in multiple expansions of Arcidae mitogenome. Although other bilaterian taxa have expansion regions in their mitochondrial genome, those in Arcidae are most extreme. us, Arcidae may provide significant insights on how bilaterian mitochondrial genomes evolve in terms of size variation, gene complement, and gene organization.

Methods

Specimen collection and sequencing

Adult specimens were obtained from populations near Qingdao (the tuandao market), Shandong Province, China. One specimen of , was collected from a local market in the Philippines. After collection, all specimens were immediately preserved at − 80 °C or in 95% ethanol. Species were identified using morphology and genetic distance of Arcidae mitochondrial . Total genomic DNA of four and one individuals was extracted from adductor muscle using the TIANamp Marine Animals DNA Kit according to the manufacturer's instructions and sequenced on an Illumina HiSeq X using 2 × 150 paired-end (PE) library. Total mRNA extraction of six

individuals was performed from adductor muscle using a Trizol protocol. All procedures were carried out on ice and quickly to avoid RNA degeneration.

e extracted mRNA was sequenced with a paired-end (PE) library with an insert size of 250 bp. e sequencing of genomic DNA and RNA was both performed by Tianjin Novogene Bioinformatics Technology Co., Ltd., China.

Mitochondrial genome assemblies and annotation

Raw data from four , one (SRX8857271) and one individuals were filtered using Trimmomatic 0.39 [77] for removal of TruSeq adapter sequence and trimming low-quality bases from the ends of each read. Clean short reads were assembled de novo using Novoplasty 4.0 [78], Mitoz v2.3 [79] with the all module and SPAdes v3.11.1 [80] with k-mer of 21, 33, 55, 77 and the –careful flag, respectively.

en, assembly results were searched using BLAST [47] against a nucleotide database constructed from the complete mitogenome of (AB729113) to find the mitochondrial contigs. Some partial mtDNA contigs were recovered in the assembly results of SPAdes. In order to bridge contigs together into a single contig, Price1.0 [81] was applied to extend and join these partial contigs with default settings by iteratively adding sequence reads to the contig ends. Finally, mitogenomes from di erent assemblers were assessed using Quast 5.0 [82] based on genome fraction, total aligned length, duplication ratio, and level of completeness. e above programs were installed on Linux system via Bioconda [83]. All the newly sequenced mitogenome sequences have already been deposited in GenBank, and the accession numbers are listed in Table 1.

Locations of the protein-coding genes (PCGs), transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) were determined initially with the MITOS web server using the invertebrate mitochondrial genetic code and validated using MFannot [84]. ORF Finder (https://www.ncbi.nlm.nih.gov/or nder/) and BLAST were employed to examine and adjust manually gene boundaries by comparison with the published Arcidae mitochondrial genes since MITOS seems to underestimate gene length.

e visualization of mitogenomes was performed using CGView [85]. Large URs structure was defined using blastn with manual alignments. ORFs in large URs of six mitogenomes were identified with ORF Finder using the invertebrate mitochondrial genetic code.

Transcriptome analysis

Quality assessment of RNA reads from six individuals was carried out using FastQC v0.11.9 [86]. Trimmomatic 0.39 was employed for removal of adaptor sequences and low-quality positions. en HISAT2 [87] was used to align clean reads from each individual to the reference mitogenomes that had been newly assembled.

e mapped reads were sorted, indexed using Samtools [88] and were assembled using Stringtie [89] in a reference-based approach. Next, we used BLAST to search the StringTie gene sets against the database of UniProtKB/Swiss-Prot proteins [90], and identified the protein domain with PFAM [91]. TransDecoder v5.5.0 was used to predict ORFs (>20 aa) that had coding potential from assembled transcripts. e results from BLASTP and hmmscan [92] searches were used to inform the final TransDecoder prediction step.

To estimate expression levels, we summarized information on annotated mitochondrial PCGs and predicted ORFs and used Artemis [93] to make a GTF file ((4) as a reference mitogenome). FeatureCounts v2.0.1 [94] was used to count mapped reads on BAM files with the option "--primary, -t exon, -g gene_id". All other parameter values were left at their defaults. BAM files were from the HISAT2 results of three transcriptomes, which are from the same experimental batch. Finally, we used a R pipeline (Additional file 2) to estimate expression levels in TPM for each mitochondrial PCGs and ORFs. Heatmap and boxplot were made with R package pheatmap and ggplot2 [95].

Novel mitochondrial ORFs and sequence conservation

To assess the presence of novel ORFs in other Arcidae mitogenomes in which they were not annotated, and at the same time validate annotations, we used BLAST to align with E-values <0.00001 and ORF Finder to search for all possible ORFs ≥75 nucleotides long under the invertebrate mitochondrial genetic code from the 32 Arcidae mitogenomes, and then translated them into the corresponding proteins. e duplicated ORFs were classified based on BLAST results and molecular phylogeny. Because ORF protein sequences vary little within the same species, only one sequence was used for analyzing. For comparative purposes, alignments of the putative novel mitochondrial proteins in di erent Arcidae species were run with MAFFT v7.475 [96]. Finally, we used Jalview [97] to visualize the alignments. en we used a ML method implemented in KaKs_Calculatorv 2.0 [98] to estimate ratios of nonsynonymous and synonymous substitution rates (Ka/Ks) in mitochondrial PCGs and ORFs between sister pairs of . ese comparisons facilitated a better understanding of the selection pressures acting on protein coding genes and ORFs. ORF names were given according to the result of ORF Finder in .

Protein structural and functional analysis

e above-mentioned ORFs were translated and analyzed at the amino acid level. Putative transmembrane (TM) helices were identified using a variety of protein signature recognition methods implemented by the following programs: Phobius [48], TMHMM 2.0 [49] and TOPCONS [50]. Evidence of signal peptides (SPs) was sought using Phobius, TOPCONS, SignalP 4.0 [99] and TargetP 2.0 [100]. HHpred [53] were used to search for known functional sequence motifs and domains. Cofactory 1.0 [101] was used to identify Rossmann fold sequence domains and predicts their specificity for the cofactors FAD, NAD or NADP. Subcellular localizations (e.g., cell membrane, cytoplasm, nucleus, etc.) were predicted using DeepLoc-1.0 [102] and Euk-mPLoc 2.0 (Cell-PLoc 2.0 package) [103].

e following procedures were used to predict the function of ORF proteins: (1) we performed BLASTp, tBLASTx, and PSI-BLAST searches against NCBI entire non-redundant protein database (NRDB) with default parameters [47] and BLAST searches against UniProt (UniProtKB reference proteomes + Swiss-Prot) with default parameters. (2) we used HHpred for profile HMM – profile HMM comparisons, which compares HMM profiles with databases of HMMs representing proteins with annotated protein families (e.g., PFAM, SMART, CDD, COGs, KOGs) or known structure (e.g., PDB, SCOP). (2) I-TASSER, which uses a hierarchical protein structure modeling approach that is based on the secondary-structure enhanced profile–profile threading alignment to predict protein tertiary structure and function [52]. (3) PredictProtein, which predicts aspects of protein structure (secondary structure, solvent accessibility, transmembrane helices and strands, coiled-coil regions, disulfide bonds and disordered regions) and function (identification of functional regions, homology-based inference of Gene Ontology terms, comprehensive subcellular localization prediction, protein-protein binding sites, protein-polynucleotide binding sites and predictions of the e ect of point mutations [non-synonymous SNPs] on protein function) [51]. For HHpred, I-TASSER and PredictProtein, only top three results were recorded.

Phylogenetic analysis and ancestral state reconstruction

We used 12 mitochondrial PCGs to reconstruct the phylogenetic history of 35 mitogenomes: six new assemblies and twenty-six published assemblies of Arcidae and outgroup species. irty-five taxa representing 26 species were included in the phylogenetic analyses and presented in Table 199618(-.009 Tw 0 -S1 gs -.005 Tw (101)Tj 09999902221ma-6(tic)-36(, wl(tg pr)6(7pr)6(ot)r)6(e)5.-9(-)-9.8918 To(. (ide,(ions)-2g prhem-692t i

Authors' contributions
NZ, YL, KMH, LK and QL designed and coordinated the research. NZ, LK and QL collected the sample. NZ performed sequencing and bioinformatics analysis. NZ and LK analyzed the data and drafted the manuscript. NZ, YL, KMH, LK and QL participated in revising the manuscript drafts during the writing process. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China under Grant 31772414, the Fundamental Research Funds for the Central Universities under Grant 201964001, and the Ocean University of China and Auburn University foundation/integrated grant program.

Availability of data and materials
The data underlying this article are available in the article and in its Supplementary Material as well as in the GenBank Nucleotide Database and can be accessed with OM807131- OM807136. Sequencing data files are available through the NCBI Sequence Read Archive (BioProject: PRJNA809524).

Declarations

Ethics approval and consent to participate
Not applicable. Samples (deceased) were either purchased from or given by commercial fisherman.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao, China. [2]Shandong University, Qingdao, China. [3]Center for Marine Science, University of North Carolina Wilmington, Wilmington, NC 28409, USA. [4]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China.

Received: 22 June 2022 Accepted: 22 November 2022
Published online: 07 December 2022

References

1. Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M. A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. Trends Genet. 2014;30(12):555–64.
2. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The origin and diversification of mitochondria. Curr Biol. 2017;27(21):R1177–92.
3. Andersson SG, Kurland CG. Reductive evolution of resident genomes. Trends Microbiol. 1998;6(7):263–8.
4. Embley TM, Martin W. Eukaryotic evolution, changes and challenges. Nature. 2006;440(7084):623–30.
5. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999;27(8):1767–80.
6. Gissi C, Iannelli F, Pesole G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity. 2008;101(4):301–20.
7. Vallès Y, Halanych KM, Boore JL. Group II introns break new boundaries: presence in a bilaterian's genome. PLoS One. 2008;3(1):e1488.
8. Zardoya R. Recent advances in understanding mitochondrial genome diversity. F1000Research. 2020;9:270. https://doi.org/10.12688/f1000research.21490.1.
9. Lindberg WP. Phylogeny and evolution of the Mollusca. Berkeley: Univ of California Press; 2008.
10. Ghiselli F, Gomes-dos-Santos A, Adema CM, Lopes-Lima M, Sharbrough J, Boore JL. Molluscan mitochondrial genomes break the rules. Philos T R Soc B. 2021;376:1825.
11. Morton BS, Prezant RS, Wilson B. Class Bivalvia. In: Beesley PL, Ross GJB, Wells A, editors. Mollusca: the southern synthesis. Clayton: CSIRO Publishing; 1998. p. 195–234.
12. Kong L, Li Y, Kocot KM, Yang Y, Qi L, Li Q, et al. Mitogenomics reveals phylogenetic relationships of Arcoida (Mollusca, Bivalvia) and multiple independent expansions and contractions in mitochondrial genome size. Mol Phylogenet Evol. 2020;150:106857.
13. Liu Y, Kurokawa T, Sekino M, Tanabe T, Watanabe K. Complete mitochondrial DNA sequence of the ark shell Scapharca broughtonii: an ultra-large metazoan mitochondrial genome. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics. 2013;8(1):72–81.
14. Hou Y, Wu B, Liu Z, Yang A, Ren J, Zhou L, et al. Complete mitochondrial genome of ark shell Scapharca subcrenata. Mitochondrial DNA Part A. 2016;27(2):939–40.
15. Sun S, Li Q, Kong L, Yu H. Evolution of mitochondrial gene arrangements in Arcidae (Bivalvia: Arcida) and their phylogenetic implications. Mol Phylogenet Evol. 2020;150:106879.
16. Smith DR, Snyder M. Complete mitochondrial DNA sequence of the scallop Placopecten magellanicus: evidence of transposition leading to an uncharacteristically large mitochondrial genome. J Mol Evol. 2007;65(4):380–91.
17. 1)7(oSpan<</ActualISME jo)

31. Cohen P. New role for the mitochondrial peptide humanin: protective agent against chemotherapy-induced side e□ects. J Natl Cancer Inst. 2014;106:3.
32. Lee C, Yen K, Cohen P. Humanin: a harbinger of mitochondrial-derived peptides? Trends in Endocrinology & Metabolism. 2013;24(5):222–8.
33. Zhan X, Zhang S, Gu Z, Wang A. Complete mitochondrial genomes of two pearl oyster species (Bivalvia: Pteriomorphia) reveal novel gene arrangements. J Shellfish Res. 2018;37(5):1039–50.
34. Breton S, Beaupré HD, Stewart DT, Piontkivska H, Karmakar M, Bogan AE, et al. Comparative mitochondrial genomics of freshwater mussels (Bivalvia: Unionoida) with doubly uniparental inheritance of mtDNA: gender-specific open reading frames and putative origins of replication. Genetics. 2009;183(4):1575–89.
35. Breton S, Ghiselli F, Passamonti M, Milani L, Stewart DT, Hoeh WR. Evidence for a fourteenth mtDNA-encoded protein in the female-transmitted mtDNA of marine mussels (Bivalvia: Mytilidae). PLoS One. 2011;6(4):e19365.
36. Breton S, Stewart DT, Shepardson S, Trdan RJ, Bogan AE, Chapman EG, et al. Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? Mol Biol Evol. 2011;28(5):1645–59.
37. Milani L, Ghiselli F. Mitochondrial activity in gametes and transmission of viable mtDNA. Biol Direct. 2015;10(1):22.
38. Milani L, Ghiselli F, Guerra D, Breton S, Passamonti M. A comparative analysis of mitochondrial ORFans: new clues on their origin and role in species with doubly uniparental inheritance of mitochondria. Genome Biology and Evolution. 2013;5(7):1408–34.
39. Milani L, Ghiselli F, Maurizii MG, Nuzhdin SV, Passamonti M. Paternally transmitted mitochondria express a new gene of potential viral origin. Genome biology and evolution. 2014;6(2):391–405.
40. Mitchell A, Guerra D, Stewart D, Breton S. In silico analyses of mitochondrial ORFans in freshwater mussels (Bivalvia: Unionoida) provide a framework for future studies of their origin and function. BMC Genomics. 2016;17(1):597.
41. Sun S, Kong L, Yu H, Li Q. The complete mitochondrial DNA of Tegillarca granosa and comparative mitogenomic analyses of three Arcidae species. Gene. 2015;557(1):61–70.
42. Sun S, Kong L, Yu H, Li Q. The complete mitochondrial genome of Scapharca kagoshimensis (Bivalvia: Arcidae). Mitochondrial DNA. 2015;26(6):957–8.
43. Sun S, Li Q, Kong L, Yu H. Limited locomotive ability relaxed selective constraints on molluscs mitochondrial genomes. Sci Rep. 2017;7(1):10628.
44. Feng Y, Li Q, Kong L. Molecular phylogeny of Arcoidea with emphasis on Arcidae species (Bivalvia: Pteriomorphia) along the coast of China: challenges to current classification of arcoids. Mol Phylogenet Evol. 2015;85:189–96.
45. Wu X, Xu X, Yu Z, Kong X. 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. 2009;2(1):69.
46. Wu X, Li X, Li L, Xu X, Xia J, Yu Z. New features of Asian Crassostrea oyster mitochondrial genomes: a novel alloacceptor tRNA gene recruitment and two novel ORFs. Gene. 2012;507(2):112–8.
47. Altschul SF, Madden TL, Schä□er AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
48. Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007;35(suppl 2):W429–32.
49. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–80.
50. Bernsel A, Viklund H, Hennerdal A, Elofsson A. TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res. 2009;37(suppl 2):W465–8.
51. Rost B, Yachdav G, Liu J. The predictprotein server. Nucleic Acids Res. 2004;32(suppl 2):W321–6.
52. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC bioinformatics. 2008;9(1):40.
53. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33(suppl 2):W244–8.
54. Combosch DJ, Giribet G. Clarifying phylogenetic relationships and the evolutionary history of the bivalve order Arcida (Mollusca: Bivalvia: Pteriomorphia). Mol Phylogenet Evol. 2016;94:298–312.
55. Marko PB. Fossil calibration of molecular clocks and the divergence times of geminate species pairs separated by the isthmus of Panama. Mol Biol Evol. 2002;19(11):2005–21.
56. Matsumoto M. Phylogenetic analysis of the subclass Pteriomorphia (Bivalvia) from mtDNA COI sequences. Mol Phylogenet Evol. 2003;27(3):429–40.
57. Akasaki T, Nikaido M, Tsuchiya K, Segawa S, Hasegawa M, Okada N. Extensive mitochondrial gene arrangements in coleoid Cephalopoda and their phylogenetic implications. Mol Phylogenet Evol. 2006;38(3):648–58.
58. Boore JL. The complete sequence of the mitochondrial genome of Nautilus macromphalus (Mollusca: Cephalopoda). BMC Genomics. 2006;7(1):182.
59. Serb JM, Lydeard C. Complete mtDNA sequence of the north American freshwater mussel, Lampsilis ornata (Unionidae): an examination of the evolution and phylogenetic utility of mitochondrial genome organization in Bivalvia (Mollusca). Mol Biol Evol. 2003;20(11):1854–66.
60. Burger G, Gray MW, Lang BF. Mitochondrial genomes: anything goes. Trends Genet. 2003;19(12):709–16.
61. Bai C, Xin L, Rosani U, Wu B, Wang Q, Duan X, et al. Chromosomal-level assembly of the blood clam, Scapharca (Anadara) broughtonii, using long sequence reads and hi-C. GigaScience. 2019;8(7):giz067.
62. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692–702.
63. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313–26.
64. Passamonti M, Ricci A, Milani L, Ghiselli F. Mitochondrial genomes and doubly uniparental inheritance: new insights from Musculista senhousia sex-linked mitochondrial DNAs (Bivalvia Mytilidae). BMC Genomics. 2011;12(1):442.
65. Passamonti M, Scali V. Gender-associated mitochondrial DNA heteroplasmy in the venerid clam Tapes philippinarum (Mollusca Bivalvia). Curr Genet. 2001;39(2):117–24.
66. Chong RA, Mueller RL. Low metabolic rates in salamanders are correlated with weak selective constraints on mitochondrial genes. Evolution: international journal of organic. Evolution. 2013;67(3):894–9.
67. Sun S, Li Q, Kong L. Relaxation of selective constraint on the ultra-large mitochondrial genomes of Arcidae (Mollusca: Bivalvia). J Ocean Univ China. 2021;20(5):1157–66.
68. Strotz LC, Saupe EE, Kimmig J, Lieberman BS. Metabolic rates, climate and macroevolution: a case study using Neogene molluscs. Proc R Soc B. 1885;2018(285):20181292.
69. Li A, Li L, Zhang Z, Li S, Wang W, Guo X, et al. Noncoding variation and transcriptional plasticity promote thermal adaptation in oysters by altering energy metabolism. Mol Biol Evol. 2021;38(11):5144–55.
70. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome biology and evolution. 2010;2:393–409.
71. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994–1006.
72. Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 2021;22(1):120.
73. Hommelsheim CM, Frantzeskakis L, Huang M, Ülker B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep. 2014;4(1):5052.
74. Hu M, Jex AR, Campbell BE, Gasser RB. Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. Nat Protoc. 2007;2(10):2339–44.
75. Kono N, Arakawa K. Nanopore sequencing: review of potential applications in functional genomics. Develop Growth Di□er. 2019;61(5):316–26.
76. Van-Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34(9):666–81.

77. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
78. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45(4):e18.
79. Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res. 2019;47(11):e63.
80. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
81. Ruby JG, Bellare P, DeRisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3: genes, genomes. Genetics. 2013;3(5):865–80.
82. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
83. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15(7):475–6.
84. Beck N, Lang B. MFannot, organelle genome annotation websever. Montreal: Université de Montréal QC; 2010.
85. Grant JR, Stothard P. The CGView server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 2008;36(suppl 2):W181–4.
86. Andrews S. FastQC. A qual control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
87. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
88. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
89. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
90. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(D1):D204–12.
91. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32.
92. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC bioinformatics. 2010;11(1):431.
93. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944–5.
94. Liao Y, Smyth GK, Shi W. featureCounts: an e□cient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
95. Wickham H. ggplot2. Wiley interdisciplinary reviews: computational statistics. 2011;3(2):180–5.
96. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
97. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.
98. Wang S, Huang Y, Liu S, Lin Z, Zhang Y, Bao Y. Hemoglobins from Scapharca subcrenata (Bivalvia: Arcidae) likely play an bactericidal role through their peroxidase activity. Comparative biochemistry and physiology Part B, Biochemistry & molecular biology. 2021;253:110545.
99. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6.
100. Emanuelsson O, Nielsen H, Brunak S, Von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000;300(4):1005–16.
101. Geertz-Hansen HM, Blom N, Feist AM, Brunak S, Petersen TN. Cofactory: sequence-based prediction of cofactor specificity of Rossmann folds. Proteins: Structure, Function, and Bioinformatics. 2014;82(9):1819–28.
102. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33(21):3387–95.
103. Chou KC, Shen HB. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Sci. 2010;2(10):1090.
104. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–73.
105. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS One. 2016;11(10):e0163962.
106. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and e□ective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
107. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3(2):217–23.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com