Acta Biochim Biophys Sin (2009): 206–216 | © The Author 2009. Published by ABBS Editorial Office in association with Oxford University Press on behalf of the Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. DOI: 10.1093/abbs/gmp003.

The complete mitochondrial genome sequence of *Geisha distinctissima* (Hemiptera: Flatidae) and comparison with other hemipteran insects

Nan Song[1,2] and Aiping Liang[1]*

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[2]Graduate School of Chinese Academy of Sciences, Beijing 100039, China
*Correspondence address. Tel: +86-10-64807226; Fax: +86-10-64807099; E-mail: liangap@ioz.ac.cn

The complete nucleotide sequence of the mitochondrial genome (mitogenome) of *Geisha distinctissima* (Hemiptera: Flatidae) has been determined in this study. The genome is a circular molecule of 15,971 bp with a total A+T content of 75.1%. The gene content, order, and structure are consistent with the *Drosophila yakuba* genome structure and the hypothesized ancestral arthropod genome arrangement. All 13 protein-coding genes are observed to have a putative, inframe ATR methionine or ATT isoleucine codons as start signals. Canonical TAA and TAG termination codons are found in nine protein-coding genes, and the remaining four (*cox1*, *atp6*, *cox3*, and *nad4*) have incomplete termination codons. The anticodons of all transfer RNA (tRNAs) are identical to those observed in *D. yakuba* and *Philaenus spumarius*, and can be folded in the form of a typical clover-leaf structure except for $tRNA^{Ser(AGN)}$. The major non-coding region (the A + T-rich region or putative control region) between the small ribosomal subunit and the $tRNA^{Ile}$ gene includes two sets of repeat regions. The first repeat region consists of a direct 152-bp repetitive unit located near the *srRNA* gene end, and the second repeat region is composed of a direct repeat unit of 19 bp located toward $tRNA^{Ile}$ gene. Comparisons of gene variability across the order suggest that the gene content and arrangement of *G. distinctissima* mitogenome are similar to other hemipteran insects.

Keywords mitochondrial genome; A + T-rich region; *Geisha distinctissima*; Flatidae; Hemiptera

Received: October 7, 2008 Accepted: November 26, 2008

Introduction

Animal mitochondrial genomes are usually circular molecules spanning 14–19 kb that contain 13 protein-coding, 2 ribosomal RNA, and 22 transfer RNA (tRNA) genes [1]. Non-coding control elements that regulate the transcription and replication of the genome are also present in mitochondrial DNAs (mtDNAs) [1,2]. Mitochondrial genomes are very important subject for different scientific disciplines, for instance, in animal health, in comparative and evolutionary genomics, in molecular evolution, and in phylogenetics and population genetics [3]. However, current knowledge on mtDNAs is very uneven as well exemplified by sequences available in GenBank that were obtained mostly from vertebrate taxa.

Insects constitute the most species-rich class among animals with almost a million of taxa described to date [4]. Within the insects, the Heteroptera (true bugs), the Sternorrhyncha (aphids, scale bugs, whiteflies, and psyllids), and the Auchenorrhyncha (planthoppers, leafhoppers, spittlebugs, and cicadas) comprise the largest non-holometabolan insect assemblage-Hemiptera [5,6]. The distinctive piercing and sucking mouthparts of Hemiptera may have played an adaptive role in their extensive evolutionary radiation [7]. High reproductive rates and an exceptional ability to transmit diseases make hemipterans, some of the worst pests, known to agriculture. In Hemiptera, the complete mitogenome nucleotide sequences have been determined for one true bug, one spittlebug, one psyllid, one aphid, one leafhopper, and six whiteflies; but there is no one complete sequence existing for planthoppers (Hemiptera: Fulgoroidea) to date. The superfamily Fulgoroidea comprises approximately 20 described insect families; many of them produce waxy skeins, such as Flatidae and Lophopidae. The family Flatidae is one of the larger families of Fulgoroidea and contains more than 1000 species in the world. The Chinese fauna is distinctive and includes 40 species, of which many cause economic damage to plants.

Geisha distinctissima is a member of the family Flatidae, and a species of considerable economic importance. Adult females oviposit in the plant tissue of their hosts (e.g. apple and citrus), and larvae feed by sucking plant juices, resulting in a significant quantitative and qualitative loss in the production of fruits. A better understanding of the hemipteran mtDNA requires an expansion of taxon and genome samplings. We have fully sequenced the mitochondrial genome of *G. distinctissima*. The newly determined mtDNA is the first complete sequence for the family Flatidae.

Materials and Methods

Sample and DNA extraction

An adult specimen of *G. distinctissima* was collected from Zhejiang province, China. After an examination of external morphology for identification, the specimen was preserved in 100% ethanol and stored at −80°C in the Key Laboratory of Animal Evolution and Systematics, Institute of Zoology, Chinese Academy of Sciences (Beijing, China).

The muscle tissue under pronotum was homogenized in 2-ml chilled buffer (220 mM mannitol, 70 mM sucrose, 5 mM Tris, and 2 mM EDTA, pH 8.0), and centrifuged at 800 *g* to pellet the nuclei and cellular debris. After the resultant supernatant was recovered through centrifugation at 3600 *g*, 1-ml homogenizing mixture was added to the precipitate and centrifuged again to pellet the mitochondria.

The pellet was digested in the protease buffer (100 mM Tris, 40 mM Nacl, 2 mM EDTA, 10% SDS, and 5 μl of 20 mg/ml proteinase K). The solution was mixed with 250 μl of 5.3 M NaCl and centrifuged at 1400 *g*. After 560 μl of isopropanol was added into the supernatant, the mixture was stewed at −20°C for 30 min and pelleted through centrifugation. The pellets were washed with 70% ethanol and stored at −20°C. DNA was dissolved in 100 μl of double distilled water and one-tenth dilutions were used as a template in PCR.

PCR amplification, cloning, and sequencing

The genome was amplified in overlapping PCR fragments (detailed primers' information is shown in **Table 1** and **Fig. 1** for a map of the amplification fragments). Initial rounds of amplification for genome sequencing were performed using standard short genes (*cox1*, *cox2*, *nad5*, *cytb*, *lrRNA*, and *srRNA*) with sets of heterologous primers that we have developed based on aligned insect and hexapod sequences. The sequences obtained from these regions were then used to design specific primers for long PCRs that allowed us to link all of the shorter regions in the following manner: $tRNA^{Met}$→*cox1*; *cox1*→*cox2*; *cox2*→*atp6*; *atp6*→*cox3*; *cox3*→*nad5*; *nad5*→*nad4*; *nad4*→*cytb*; *cytb*→*lrRNA*; *lrRNA*→*srRNA*; *srRNA*→$tRNA^{Met}$. In this manner, we amplified the entire mitogenome.

Short PCRs were conducted using Qiagen *Taq* DNA polymerase (Qiagen, Beijing, China) with the following cycling conditions: 5 min at 94°C, followed by 30 cycles of 50 s at 94°C, 50 s at 50°C, and 1–2 min at 72°C. The final elongation step was continued for 10 min at 72°C. As for large fragments, long PCRs were performed using Qiagen Long *Taq* DNA polymerase (Qiagen) under the following cycling conditions: 2 min at 96°C, followed by 30 cycles of 10 s at 98°C, and 10 min at 68°C. The final elongation was continued for 10 min at 72°C. These PCR products were analyzed by 1.0% agarose gel electrophoresis.

PCR products of ∼1200 bp (fragments 1–7, 9–13, 16, 18–21, and 23 in **Fig. 1**) were directly sequenced after purification, but the PCR products of 1.2–2.1 kb (fragments 8, 14, 15, 17, and 22 in **Fig. 1**) were cloned into pBS-T Easy vector (Qiagen) and the resultant plasmid DNA was isolated using the TIANprp Midi Plasmid Kit (Qiagen). For each larger PCR product, at least two independent clones were sequenced to ensure that we obtained the true sequence. As those PCR fragments of the cloned DNA are sequenced, internal primers were designed to further sequence into the fragments. DNA sequencing was performed using the BigDye Terminator Cycle Sequencing Kit and the ABI 3730XL Genetic Analyzer (PE Applied Biosystems, San Francisco, CA, USA). All fragments were sequenced from both strands.

Sequence assembly, annotation, and analysis

Raw sequence files were proof-read and aligned into contigs in BioEdit version 7.0.5.3 [8]. Contig sequences were checked for ambiguous base calls and only non-ambiguous regions were used for annotation. Sequences alignment, genome assemblage, and nucleotide composition calculations were all conducted with Mega 4 [9]. The locations of protein-coding genes and rRNA genes were identified by determining sequence similarity with other insects, whereas tRNA genes were identified using the tRNAscan-SE server [10]. The tRNAs not found by tRNAscan-SE were identified by comparing the regions coding these tRNAs in other insects. Potential secondary structure folds in the mitogenome were predicted using Mfold v. 3.2 (http://www.bioinfo.rpi.edu/applications/mfold/) [11]. Sequence

Table 1 Primers used in this study

Upstream primers	Sequence (5′→3′)	Downstream primers	Sequence (5′→3′)
F01	AATTGGTGGTTTTGGAAATTG	R01	GGTAATCAGAGTATCGACG
F02	TCTAATATGGCAGATTAGTGCA	R02	ACTATTAGATGGTTTAAGAG
F03	AGAGGTATATCACTGTTAATGA	R03	TTAGGTTGAGATGGTTTAGG
F04	CTCATACTGATGAAATTTTGGTTC	R04	TTCTACTGGTCGTGCTCCAATTCA
F05	CCGGTCTGAACTCAGATCAT	R05	ATTTATTGTACCTTTTGTATCAG
F06	GTAAAYCTACTTTGTTACGACTT	R06	GTGCCAGCAAYCGCGGTTATAC
F07	CATGAATAGATTATTTATAATAACAT	R07	TGGTAAAAATCCTATTATGGG
F08	AATAAAGCTAATAGGTTCATACCC	R08	TTTATTCGTGGAAATGCTATGTC
F09	GTTAAATAAACTAGTAACCTTCAAA	R09	GCTCGTGTATCAACGTCTATACC
F10	TGATTTTTTGGTCATCCAGAAGT	R10	TATTCATATCTTCAATATCATTGATG
F11	CAATGCTCAGAAATTTGTGG	R11	ATGACCTGCAATTATATTAGC
F12	ATTAAATGATAACAAGATTATTTTC	R12	CGTACTACGTCTCGTCATCATAT
F13	GTTGACTATAGCCCTTGACC	R13	TCAATTTTRTCATTAACAGTGA
F14	CCATTTGAATGTGGRTTTGATCC	R14	TTAAGGCTTTATTATTTATATGTGC
F15	AAACGGAAACTGAGCACTTTTAGT	R15	TGAGGTTATCARCCTGAACG
F16	CCAGAAGAACATAANCCATG	R16	TGAGGTTATCARCCTGAACG
F17	ATAGTAGGTCCTTCTACATGAGC	R17	TATCAATAGCGAATCCTCCTCA
F18	CGTTCAGGTTGATACCCCCA	R18	TTTCGTTTGAGGCCACTC
F19	TCCATATTCAACCAGAATGATA	R19	TTTGTTTCCTGGTCTTGGG
F20	AGGAAAGGAACCACGAACCCA	R20	ATACCTTAGGGATAACAGCGTGA
F21	CCTTTGTACAGTTAAAATACTGC	R21	AATTATGTACATATCGCCCTTC
F22	ATAATAGGGTATCTAATCCTAGT	R22	ACCTTTATAAATGGGGTATGAACC
F23	AAAGCTAACCCATGAGGT	R23	AAACAATAATGCATGAATAG

F, forward; R, reverse.

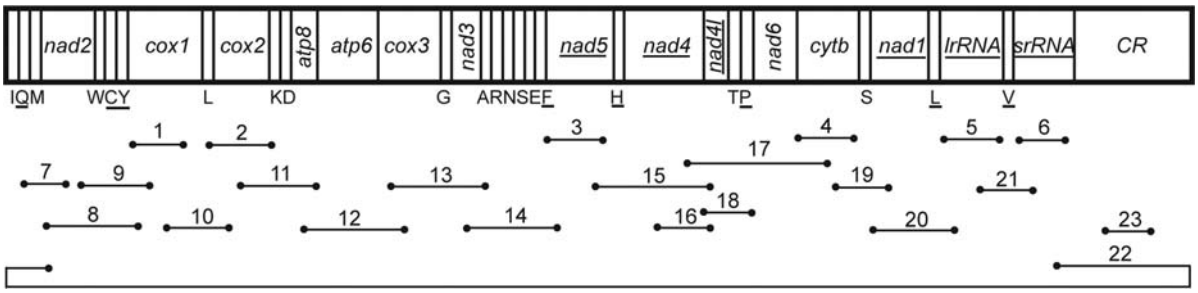

Fig. 1 Schematic representation of amplification strategy employed for the *G. distinctissima* mitochondrial genome Lines below the linearized genome map represent the amplification products.

data are available from NCBI (http://www.ncbi.nlm.nih.gov/) under accession number FJ230961.

Results

Genome structure, organization, and composition

The completed sequence of the *G. distinctissima* mitogenome is 15,971 bp in length. Within this mitogenome, each of the 37 genes typically found in most insects is discovered: 22 transfer RNA genes, 13 protein-coding genes, and 2 ribosomal RNA genes. Besides, there is one major non-coding region [the A + T-rich region or putative control region (CR)] between the small ribosomal subunit and the $tRNA^{Ile}$ gene (**Table 2**). The structure of the mitogenome of *G. distinctissima* and amplified regions are shown in **Fig. 1**.

Table 2 Summary of the mitochondrial genes of *G. distinctissima*

Gene	Direction	Span (bp)	Size (bp)	Anticodon	Start	Stop
tRNA^{Ile}	F	1–63	63	GAT (31–33)		
tRNA^{Gln}	R	63–131	69	TTG (99–101)		
tRNA^{Met}	F	131–196	66	CAT (162–164)		
nad2	F	197–1162	966		ATA	TAA
tRNA^{Trp}	F	1161–1224	64	TCA (1191–1193)		
tRNA^{Cys}	R	1217–1279	63	GCA (1248–1250)		
tRNA^{Tyr}	R	1283–1344	61	GTA (1312–1314)		
cox1	F	1351–2884	1534		ATG	T-tRNA
tRNA$^{Leu(UUR)}$	F	2885–2949	65	TAA (2914–2916)		
cox2	F	2951–3637	687		ATA	TAA
tRNA^{Lys}	F	3640–3706	67	CTT (3669–3671)		
tRNA^{Asp}	F	3707– 3771	65	GTC (3737–3739)		
atp8	F	3772–3924	153		ATT	TAA
atp6	F	3918–4569	652		ATG	T-cox3
cox3	F	4570–5350	781		ATG	T-tRNA
tRNA^{Gly}	F	5351–5412	62	TCC (5381–5383)		
nad3	F	5413–5760	348		ATT	TAA
tRNA^{Ala}	R	5761–5825	65	TGC (5790–5792)		
tRNA^{Arg}	F	5826–5886	61	TCG (5853–5855)		
tRNA^{Asn}	F	5892–5953	62	GTT (5922–5924)		
tRNA$^{Ser(AGN)}$	F	5954–6014	61	GCT (5982–5984)		
tRNA^{Glu}	F	6017–6078	62	TTC (6047–6049)		
tRNA^{Phe}	R	6082–6145	64	GAA (6110–6112)		
nad5	R	6136–7782	1647		ATG	TAG
tRNA^{His}	R	7792–7858	67	GTG (7827–7829)		
nad4	R	7859–9179	1321		ATG	T-tRNA
nad4l	R	9188–9469	282		ATG	TAG
tRNA^{Thr}	F	9472–9532	61	TGT (9502–9504)		
tRNA^{Pro}	R	9536–9597	62	TGG (9565–9567)		
nad6	F	9588–10091	504		ATA	TAA
Cytb	F	10,084–11,205	1112		ATG	TAA
tRNA$^{Ser(UCN)}$	F	11,206–11,266	61	TGA (11,233–11,235)		
nad1	R	11,273–12,205	933		ATG	TAA
tRNA$^{Leu(CUN)}$	R	12,207–12,268	62	TAG (12,236–12,238)		
lrRNA	R	12,269–13,466	1198			
tRNA^{Val}	R	13,467–13,540	74	TAC (13,503–13,505)		
srRNA	R	13,541–14,269	729			
A + T rich region		14,270–15,971	1702			
Repeat region 1		14,308–14,932	625			
Repeat region 2		15,726–15,841	116			

The overall organization of the *G. distinctissima* mitogenome is very compact, with only 51 nucleotides dispersed in 13 intergenic spacers and ranged in size between 1 and 9 bp. The longest spacer sequence (9 bp) is located between *nad5* and *tRNA^{His}*. The contiguous genes overlap in a total of 51 bp at nine locations ranging from 1 to 10 bp, with the two largest measuring 10 bp located between *tRNA^{Phe}* and *nad5*, and between *tRNA^{Pro}* and *nad6*.

The nucleotide composition bias of the *G. distinctissima* mitogenome is 75.1% A + T. This is similar to the

nucleotide composition biases found in other hemipteran species that range from 69.5% in *Triatoma dimidiata* to 86.3% in *Aleurodicus dugesii* (**Table 3**). Within protein-coding genes of *G. distinctissima*, the A+T composition is 73.5% and 77.0% for total tRNAs, 78.8% for rRNA genes, and 79.5% for the A + T-rich region. The nucleotide skew statistics [12] for the whole mitogenome (measured on the majority strand) reveal that *G. distinctissima* is A- and C-skewed. Among other hemipterans, *T. dimidiata* has a similar A- and C-skew values to *G. distinctissima*, but the two whiteflies (*A. dugesii* and *Trialeurodes vaporariorum*) evidence an opposite bias, and the remaining species lack significant A-skew and have a moderate C-skew. Dividing the total protein-coding genes to those encoded on the majority strand (*cox1*, *cox2*, *cox3*, *cytb*, *atp6*, *atp8*, *nad2*, *nad3*, and *nad6*) and those encoded on the minority strand (*nad1*, *nad4*, *nad4l*, and *nad5*) shows that the mitogenome nucleotide compositions are due to a strong T-biased in the minority strand genes of *G. distinctissima* (A-skew = −0.5257), and the majority strand genes strongly C-skewed.

Protein-coding genes

The standard animal mitogenome encodes 13 protein-coding genes that are subunits of four of the five mitochondrial membrane-associated protein complexes involved in oxidative phosphorylation. The genes are *NADH dehydrogenase subunit 1* (*nad1*), *nad2*, *nad3*, *nad4*, *nad4l*, *nad5*, and *nad6* from complex I (NADH-ubiquinol oxidoreductase), *cytochrome b* (*cytb*) from complex III (ubiquinone-cytochrome-c-oxidoreductase), *cytochrome c oxidase subunit 1* (*coxl*), *cox2*, and *cox3* from complex IV (cytochrome c oxidase), and *ATP synthase FO subunit 6* (*atp6*) and *atp8* from complex V (ATP synthase). The mtDNA of *G. distinctissima* contains the full set of protein-coding genes usually present in animal mtDNA. A summary of the mitochondrial genes of *G. distinctissima* is given in **Table 2**.

All 13 protein-coding genes are found to have a putative, inframe ATR methionine or ATT isoleucine codons as start signals. Eight start codons are coded by ATG (*cox1*, *atp6*, *cox3*, *nad5*, *nad4*, *nad4l*, *cytb*, and *nad1*), three by ATA (*nad2*, *cox2*, and *nad6*), and two by ATT (*atp8* and *nad3*). Canonical TAA and TAG termination codons are found in seven (*nad2*, *cox2*, *atp8*, *nad3*, *nad6*, *cytb*, and *nad1*) and two (*nad5* and *nad4l*) protein-coding genes, respectively. The remaining four have incomplete termination codons.

Codon usage was calculated for all protein-coding genes (**Table 4**). Leucine, phenylalanine, isoleucine, and methionine are the four most frequently used amino acids,

Table 3 Mitochondrial genome comparisons among hemipteran species

Species	Whole genome				PCGs		*lrRNA*		*srRNA*		A + T rich region	
	Size (bp)	AT%	A-skew	C-skew	No. of codons[a]	AT%	Size (bp)	AT%	Size (bp)	AT%	Size (bp)	AT%
Geisha distinctissima	15,971	75.1	0.2650	0.2691	3633	73.5	1198	79.5	729	77.7	1702	79.5
Philaenus spumarius	16,324	77.0	0.0623	0.0783	3676	76.0	1245	79.3	754	77.9	1835	78.9
Homalodisca coagulata	15,304	78.4	0.0969	0.1204	3645	77.1	1201	80.8	728	77.1	1034	88.1
Triatoma dimidiata	17,009	69.5	0.1683	0.2656	3689	68.8	1270	74.5	781	70.9	2158	66.0
Pachypsylla venusta	14,711	75.0	0.0693	0.2560	3630	73.8	1148	78.8	765	76.4	597	83.9
Schizaphis graminum	15,721	84.0	0.0667	0.2547	3644	83.2	1259	85.4	766	84.1	636	84.9
Aleurodicus dugesii	15,723	86.3	−0.0806	−0.1913	3608	84.7	1188	88.4	730	87.6	1576	92.7
Trialeurodes vaporariorum	18,414	72.3	−0.1824	−0.2004	3599	68.0	1211	77.9	744	76.7	3729	81.3

PCGs, protein-coding genes.
[a]Termination codons were excluded in total codon count.

Table 4 Codon usage table for *G. distinctissima* mitochondrial DNA

a.a./percentage (%)	Codon	Number	a.a./percentage (%)	Codon	Number
Ala/2.67[a]	GCG	2	Pro/3.63	CCG	3
	GCA	45		CCA	67
	GCT	36		CCT	42
	GCC	14		CCC	20
Cys/1.40	TGT	45	Ser/7.32	TCG	7
	TGC	6		TCA	137
Asp/1.87	GAT	48		TCT	98
	GAC	20		TCC	24
Glu/2.29	GAG	15	Gly/5.37	GGG	38
	GAA	68		GGA	80
Phe/12.17	TTT	391		GGT	70
	TTC	51		GGC	7
Ser/2.84	AGG	8	Arg/1.38	CGG	3
	AGA	64		CGA	23
	AGT	26		CGT	19
	AGC	5		CGC	5
His/1.90	CAT	48	Thr/5.56	ACG	7
	CAC	21		ACA	133
Ile/9.50	ATT	282		ACT	33
	ATC	63		ACC	29
Lys/4.07	AAG	22	Val/4.68	GTG	13
	AAA	126		GTA	62
Leu/8.01	TTG	62		GTT	85
	TTA	229		GTC	10
Leu/4.46	CTG	13	Gln/1.54	CAG	7
	CTA	74		CAA	49
	CTT	68	Trp/2.70	TGG	22
	CTC	7		TGA	76
Met/8.37	ATG	43	Tyr/4.16	TAT	116
	ATA	261		TAC	35
Asn/4.13	AAT	117			
	AAC	33			

[a]The percentages under each amino acid three letter codes are the percentages of the amino acids found among all of the proteins (3633 a.a.). This analysis excludes stop codons.

accounting for 12.47%, 12.17%, 9.50%, and 8.37%, respectively, in the *G. distinctissima* mitochondrial proteins.

Transfer RNA and ribosomal RNA genes

The standard 22 tRNA genes are present in the *G. distinctissima* mitogenome. All tRNAs have the typical clover-leaf structure of mitochondrial tRNAs except for $tRNA^{Ser(AGN)}$, in which the dihydrouracil arm forms a simple loop (**Fig. 2**); this is also the case in several metazoan mtDNAs, including some hemipteran insects [13,14]. The *G. distinctissima* tRNA genes range in size from 61 to 74 bp, which are well within the observed range for other insects. The anticodons of the *G. distinctissima* tRNAs are identical to those in *Drosophila yakuba* and *Philaenus spumarius* [13,15]. A total of 26 unmatched base pairs occur in 15 tRNAs of *G. distinctissima*. Of these, 11 are G-U pairs, and the remaining 15 are U-U, A-A, A-C, and A-G mismatches. Similarly, the total number of mismatches in 22 tRNA genes found in *P. spumarius* is 31.

The two genes encoding the large and the small ribosomal subunits are located between $tRNA^{Leu(CUN)}$ *and*

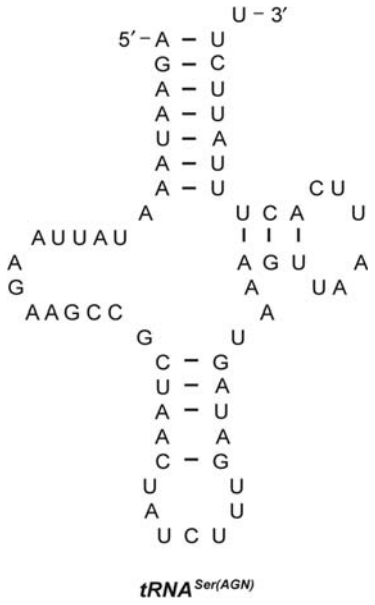

Fig. 2 Inferred secondary structure of $tRNA^{Ser(AGN)}$ of *G. distinctissima* The tRNA is labeled with the abbreviation of corresponding amino acid. Nucleotide sequences are from 5′ to 3′ as indicated for $tRNA^{Ser(AGN)}$. Watson–Crick base pairs designated by '–'.

$tRNA^{Val}$, and between $tRNA^{Val}$ and the A + T-rich region. The length of *G. distinctissima lrRNA* and *srRNA* were, respectively, determined to be 1198 and 729 bp, which were based on the location of neighboring tRNAs and a comparison with other related sequences and structures. These lengths are shorter than those of *P. spumarius* and *T. dimidiata* (1245 bp *lrRNA* and 754 bp *srRNA*; 1270 bp *lrRNA* and 781 bp *srRNA*).

A + T-rich or CR

The principal A + T-rich region (or CR) of the mitogenome is a low G+C content region usually following *srRNA*. The low G+C region usually has stretches of 'T's or 'A's as well as multiples of the sequence 'TA'. Another feature of this region may be inverted and direct repeats. The A + T-rich region of the *G. distinctissima* mitogenome was determined to be 1702 bp in length, which can be divided into five parts [**Fig. 3(A)**]: (i) a 38 bp lead sequence that bordered by small ribosomal subunit, of which the G+C content (29.0%) is higher than the whole mitogenome; (ii) the first 625-bp repeat region is immediately after the lead sequence, and this region consists of five tandem repeat units; (iii) a [TA(A)]n-like stretch including several poly-A stretch in the 3′-end, which is heavily biased toward A + T (87.6%); (iv) the second tandem repeat region, which is composed of a 19-bp repeat unit 'TGAAAAAT CAAAAAATTGA' on the majority strand; (v) a poly-T stretch near $tRNA^{Ile}$ on the minority strand, which may be involved in transcriptional control or may be the site for initiation of replication [16,17].

Discussion

The size (15,971 bp) of the whole mitogenome of *G. distinctissima* is comparable to that reported for other hemipterans, which range from 14,496 bp in *Neomaskellia andropogonis* to 18,414 bp in *T. vaporariorum* [18]. Genome arrangement of the *G. distinctissima* mitogenome is the same as the ancestral insect [1] and is identical to the other published hemipteran mitogenome (e.g. *P. spumarius*) [13].

Similar to *G. distinctissima*, other hemipteran species also have intergenic spacer sequences in the mitogenome. For example, *T. dimidiata* contains a 314-bp-long spacer sequence, which has a possibility to code for an unknown gene because the complementary strand has a complete start methionine and stop codon [14]. Also, *Schizaphis graminum* contains a 50-bp-long intergenic spacer between *nad5* and $tRNA^{His}$ [18]. Excluding these, the size of intergenic spacer reported in other insects is usually <50 bp. In most cases, the intergenic spacer sequences consist of only 1 or 2 bp. As for overlapping regions in the contiguous genes in other hemipterans, the total sizes range from 29 bp in *S. graminum* to 99 bp in *T. dimidiata* [14,18].

The problematic translational start of the *cox1* locus has been extensively discussed in several arthropod species including insects [19,20]. It was postulated to be of tetranucleotide (ATAA, TTAA, and ATTA) or hexanucleotide (ATTTAA) nature. Whereas hemipteran species do not share this feature, for instance, *G. distinctissima*, *P. spumarius*, *T. dimidiata*, and *Pachypsylla venusta* all use typical methionine (ATG) as the first amino acid for *cox1* [13,14,18]. This appears to suggest that such abnormalities are taxon-specific [21].

Incomplete termination codons (T or TA) are commonly found in many metazoan mitogenomes including that of insects [22,23]. The common interpretation of this phenomenon is that TAA termini are created via post-transcriptional polyadenylation [24]. Accordingly, partial stop codons have been observed in other hemipteran insects' mitogenomes. There seems to be a high degree of conservation of incomplete stop codons across the Hemiptera. For instance, *cox1*, *cox3*, and *nad4* genes all have incomplete termination codons (T) as stop signals

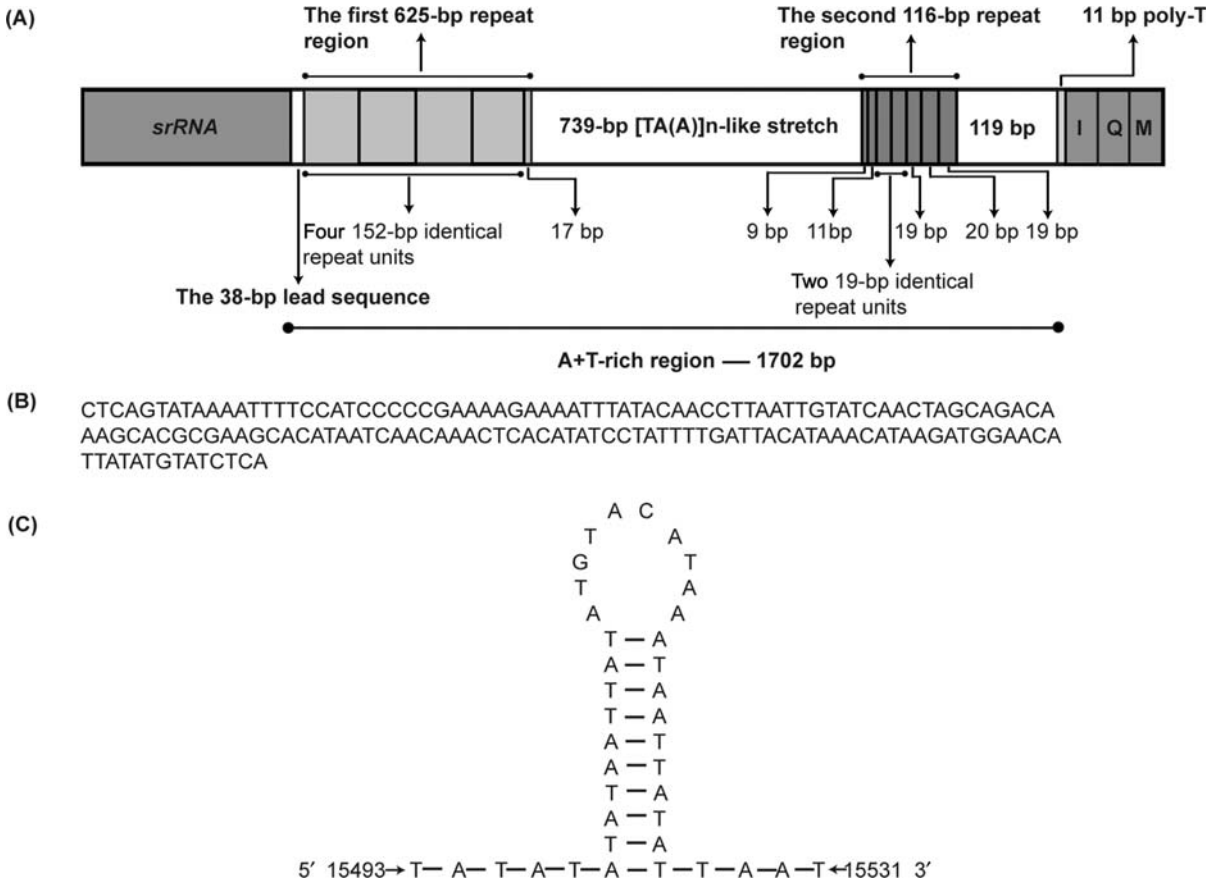

Fig. 3 The A + T-rich region of the *G. distinctissima* mitogenome (**A**) The structural organization of the A + T-rich region of the *G. distinctissima* mitogenome. The A + T-rich region flanking genes *srRNA*, $tRNA^{Ile}$ (I), $tRNA^{Gln}$ (Q), and $tRNA^{Met}$ (M) are represented in gray boxes. (**B**) The sequence of the repeat unit in the first repeat region. (**C**) The potential stem-and-loop structure in A + T-rich region of *G. distinctissima*

in *P. spumarius*, *T. dimidiata*, *P. venusta*, and *Homalodisca coagulata* [13,14,18].

In the mitogenome of *G. distinctissima*, six of the protein-coding genes (*nad2*, *cox1*, *cox2*, *nad3*, *nad5*, and *nad1*) are flanked by tRNA genes on both the 5′- and 3′-ends, six other protein-coding genes (*atp8*, *cox3*, *nad4l*, *nad4*, *nad6*, and *cytb*) are flanked on one side by a tRNA and on the other side by a protein-coding gene, and one (*atp6*) is flanked by protein-coding genes on both sides (**Table 2**). Among the latter seven protein-coding genes, four adjoin another protein-coding gene at their 3′-end: *atp8*, *atp6*, *nad4l*, and *nad6*, which are arranged as *atp8-atp6*, *atp6-cox3*, *nad4l-nad4*, and *nad6-cytb*. These are the two sets of overlapping genes (*atp8-atp6* and *nad6-cytb*), one set of abutting genes (*atp6-cox3*), and one set of interspaced genes (*nad4l-nad4*). An interesting aspect of this study is the finding that there is a 7-bp identical overlapping sequence at the junction of *atp8-atp6* genes among several hemipteran species (**Fig. 4**). The reason for this phenomenon may be that these sequences are located at the same site of the *atp8-atp6* gene cluster, and the *atp8* genes of these species have the similar transcriptional mechanism [25]. It has been proposed that the secondary structure of the transcribed polycistronic mRNA may facilitate cleavage between the proteins [15]. As for the *atp8-atp6* genes, because they are translated from a single mRNA, there is no cleavage of the mRNA for expression [25]. The same mechanism is applied to the *nad4l-nad4* genes [25]. For the other two sets of adjacent protein-coding genes (*atp6-cox3* and *nad6-cytb*), we have detected the potential secondary structures forming at the 3′-end, which may act as signals for the cleavage of the polycistronic primary transcript [15]. The similar situations have been observed at the junctions of the same protein-coding genes in other hemipteran mitogenomes (e.g. in *P. spumarius* and *T. dimidiata*) [13,14] (**Fig. 5**).

In the *G. distinctissima* mitochondrial proteins, leucine, phenylalanine, isoleucine, and methionine are the four most frequently used amino acids. Similarly,

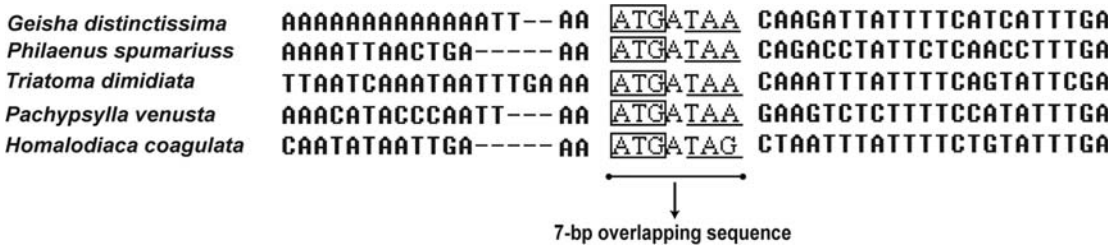

Fig. 4 Alignments of the identical 7-bp overlapping sequences at the junction of *atp8-atp6* genes among several hemipteran species Translation initiation codons for *atp6* genes are boxed, and termination codons for *atp8* genes are underlined. '−' represents inferred gaps.

these amino acids are also frequently utilized in other hemipteran insects (for example, in *P. spumarius* and *T. dimidiata*) [13,14]. Analysis of the codon usage reveals that codons with A or T in the third position are most prevalent in the mitogenome of *G. distinctissima*. The four most frequently used codons, TTA (leucine), TTT (phenylalanine), ATT (isoleucine), and ATA (methionine), are all composed wholly of T and/or A (**Table 4**). This implies that the amino acid composition is affected to a similar degree by the AT mutational bias. The length and amino acid composition of the 13 proteins of *G. distinctissima* (**Tables 2** and **4**) are similar to those of *T. dimidiata* [14].

The presence of varying copy numbers of tandemly repeated elements was reported as one of the characteristics of the insect A + T-rich region [26]. In the case of *P. spumarius*, the 1834-bp A + T-rich region harbors two repeat regions, and the two different sets of repeat units were separated by a non-repetitive sequence [13]. The 1575-bp A + T-rich region of whitefly *A. dugesii* includes five 163-bp repeat units [18]. One of the most extreme cases is the 2165-bp-long *T. dimidiata* A + T-rich region, which harbors eight tandem repeats composed of one copy of an 82-bp element, five copies of a 140-bp element, and two copies of a 173-bp element [14]. The A + T-rich region of *G. distinctissima* is more similar to that of *P. spumarius*, in that it also includes two repeat regions at both ends. The first repeat region is situated on the *srRNA* gene side of the A + T-rich region. This repeat region consists of four identical 152-bp repeat units [**Fig. 3(B)**], and the fifth repeat unit is only the copy of the first 17 nucleotides of the 152-bp repeat unit. This region is less biased toward A+T (68.7%) than the mitochondrial genome as a whole (75.1%).

After the first repeat region, there is a 793-bp A + T (87.8%) rich region, which includes many [TA(A)]n-like stretches. The stem-and-loop structure in the A + T-rich region was suggested as the site of the initiation of secondary strand synthesis in *Drosophila* [27]. Although the primary sequences of the secondary structure were found to be highly divergent, the flanking sequence of the structure was suggested to be highly conserved among some insects, possessing the consensus 'TATA' sequences at the 5′-end and 'GAA(A)T' at the 3′-end [26,28]. In the *G. distinctissima* A + T-rich region, several DNA segments have the potential to form stem-and-loop structures. These structures are formed by stems with perfect matches of varying nucleotide pairs and loops of various sizes. Among these, a single stem-and-loop structure harbors the conserved 5′-flanking 'TATA', but we did not find 'G (A) nT' in the 3′-end [**Fig. 3(C)**]. As well, the A + T-rich region of some other insects (for example, *P. spumariu* and *Coreana raphaelis*) while exhibiting the potential to form stable secondary structures did not show conservation of flanking motifs [13,29]. Thus, it seems that the immediately flanking sequence may exist as a different form, or such a sequence may not be universally conserved in insects.

The *G. distinctissima* A + T-rich region also contains a second repeat sequence of 116 bp in length. This sequence comprises of seven tandem repeat units. Among them, the first two repeat units are only the fragments of repeat unit 'TGAAAAATCAAAAAATTGA'. The third and fourth contain complete repeat unit. The fifth has two transversions that are T to C at position 17 and G to A at position 18. Besides, repeat unit 6 has an insertion of T at position 14 and a transversion of T to C at position 17, and repeat unit 7 also has a T to C transversion near the 3′-end. The whole sequence is strongly biased toward A+T (82.8%).

Conclusions

The mitochondrial genome of *G. distinctissima* is the first sequenced mtDNA for a representative of the Flatidae, a family that has considerable economic importance. The size, nucleotide composition, and genome arrangement are typical of a 'standard' insect mitogenome, such as the *D. yakuba* mitogenome [15].

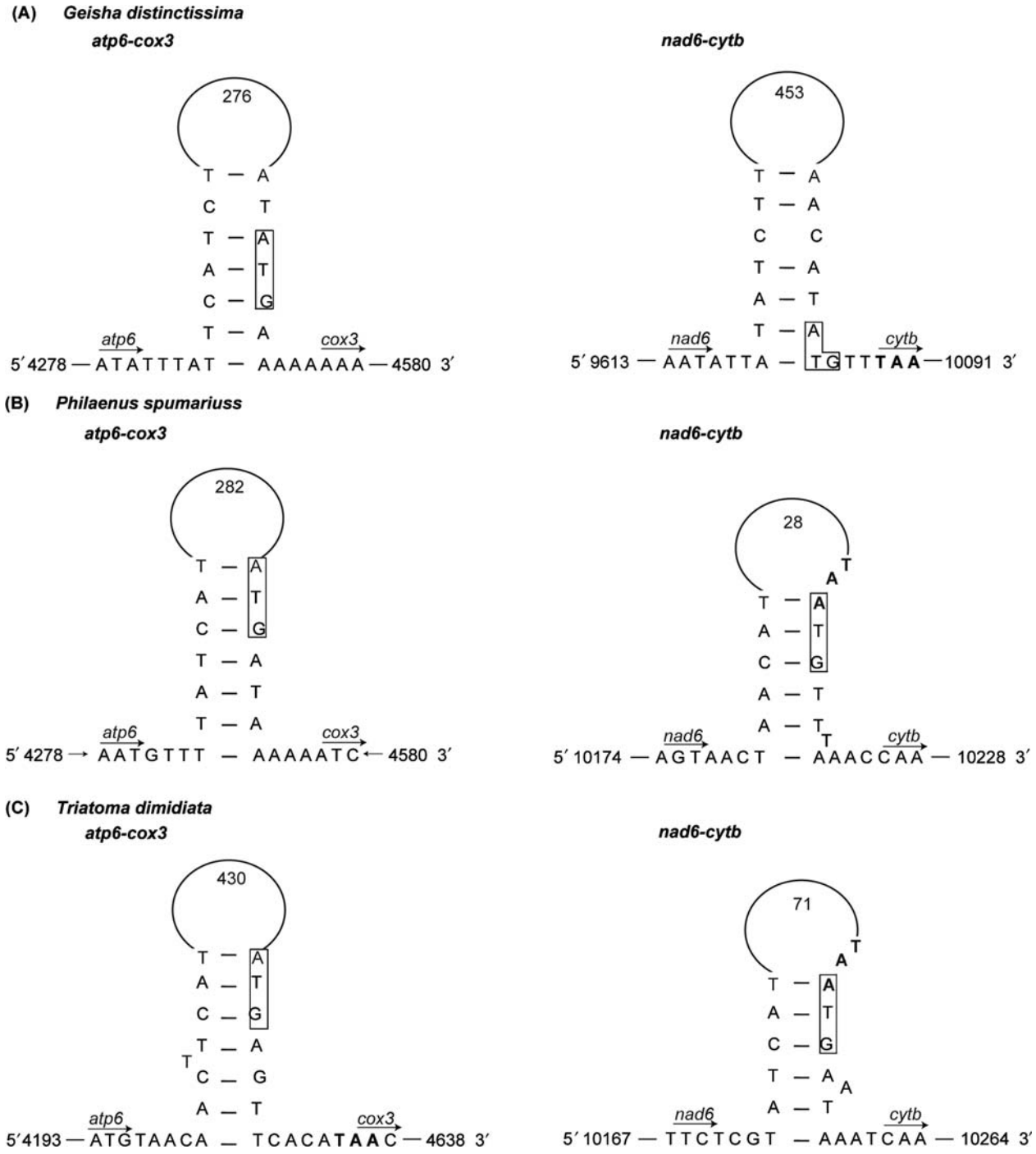

Fig. 5 Potential stem-and-loop configurations of the nucleotides of the junctional sequences of *atp6-cox3* and *nad6-cytb* of *G. distinctissima* (A), *P. spumariuss* (B), and *T. dimidiata* (C) Arrows indicate the direction of transcription. Nucleotide positions of stem-and-loop structure are indicated at each end site of stem-and-loop structure with respect to the species' mitogenome. Translation initiation codons are boxed, and termination codons (or incomplete termination codon) are bold-faced. The number of nucleotides in each loop is shown.

The 15,971 bp *G. distinctissima* mitogenome contains 75.1% A + T. Nucleotide biases are observed to be strand-specific and positional in nature. The initiation codon of the *cox1* gene appears to be ATG canonical start codons, and other hemipterans share this feature. Sequence analysis revealed the presence of two different types of tandem repeats in the *G. distinctissima* A + T-rich region, and this is similar to that of other hemipteran species, for example, *P. spumarius*. The first repeat region consists of a direct 152-bp repetitive unit

repeated four times, which is located near the *srRNA* gene end (position 38 bp from the 5′-end of the A + T-rich region). The second repeat region is composed of a direct unit of 19-bp located toward $tRNA^{Ile}$ gene.

Acknowledgement

We thank Dr Chuan Ma (Institute of Zoology, Chinese Academy of Sciences, Beijing, China) for technical assistance and advice on the experiments.

Funding

The work on which this paper is based was supported by the National Natural Science Foundation of China (grant number 30530110) and the National Science Fund for Fostering Talents in Basic Research (Special subjects in animal taxonomy, NSFC-J0630964/J0109), both awarded to APL.

References

1 Boore JL. Animal mitochondrial genomes. Nucleic Acids Res 1999, 27: 1767–1780.
2 Taanman JW. The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta 1999, 1410: 103–123.
3 Salvato P, Simonato M, Battisti A and Negrisolo E. The complete mitochondrial genome of the bag-shelter moth *Ochrogaster lunifer* (Lepidoptera, Notodontidae). BMC Genomics 2008, 9: 331.
4 Resh VH and Cardé RG. Insecta, Overview. In: Resh VH and CArdé RG eds. Encyclopedia of Insects. Burlington MA, USA: Academic Press, 2003, 564–566.
5 Kristensen NP. Phylogeny of extant hexapods. In: C.S.I.R. Organization ed. The Insects of Australia, a Textbook for Students and Research Workers, 2nd edn. Victoria: Melbourne University Press, 1991, 125–142.
6 Carpenter FM. Treatise on Invertebrate Paleontology. Vol 3, Superclass Hexapoda. Boulder, Colorado and Lawrence, Kansas: The Geological Society of America and the University of Kansas, 1992.
7 Goodchild AJP. Evolution of the alimentary canal in the Hemiptera. Biol Rev 1966, 41: 97–140.
8 Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 1999, 41: 95–98.
9 Tamura K, Dudley J, Nei M and Kumar S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 2007, 24: 1596–1599.
10 Lowe TD and Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997, 25: 955–964.
11 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 2003, 31: 3406–3415.
12 Perna NT and Kocher TD. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol 1995, 41: 353–359.
13 Stewart JB and Beckenbach AT. Insect mitochondrial genomics: the complete mitochondrial genome sequence of the meadow spittlebug *Philaenus spumariuss* (Hemiptera: Auchenorrhyncha: Cercopoidae). Genome 2005, 48: 46–54.
14 Dotson EM and Beard CB. Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*. Insect Mol Biol 2001, 10: 205–215.
15 Clary DO and Wolstenholme DR. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J Mol Evol 1985, 22: 252–271.
16 Lewis DL, Farr CL, Farquhar AL and Kaguni LS. Sequence, organization, and evolution of the A+T region of *Drosophila melanogaster* mitochondrial DNA. Mol Biol Evol 1994, 11: 523–538.
17 Zhang DX, Szymura JM and Hewitt GM. Evolution and structural conservation of the control region of insect mitochondrial DNA. J Mol Evol 1995, 40: 382–391.
18 Thao ML, Baumann L and Baumann P. Organization of the mitochondrial genomes of whiteflies, aphids, and psyllids (Hemiptera, Sternorrhyncha). BMC Evol Biol 2004, 4: 25.
19 Caterino MS and Sperling FA. Papilio phylogeny based on mitochondrial cytochrome oxidase I and II genes. Mol Phylogenet Evol 1999, 11: 122–137.
20 Wilson K, Cahill V, Ballment E and Benzie J. The complete sequence of the mitochondrial genome of the crustacean Penaeus mondon: are malacostracan crustaceans more closely related to insects than to branchiopods? Mol Biol Evol 2000, 17: 863–874.
21 Cha SY, Yoon HJ, Lee EM, Yoon MH, Hwang JS, Jin BR and Han YS, *et al*. The complete nucleotide sequence and gene organization of the mitochondrial genome of the bumblebee, *Bombus ignitus* (Hymenoptera: Apidae). Gene 2007, 392: 206–220.
22 Bae JS, Kim I, Sohn HD and Jin BR. The mitochondrial genome of the firefly, Pyrocoelia rufa: complete DNA sequence, genome organization, and phylogenetic analysis with other insects. Mol Phylogenet Evol 2004, 32: 978–985.
23 Cantatore P, Roberti M, Rainaldi G, Gadaleta MN and Saccone C. The complete nucleotide sequence, gene organization, and genetic code of mitochondrial genome of Paracentrotus lividus. J Biol Chem 1989, 264: 10965–10975.
24 Ojala D, Montoya J and Attardi G. tRNA punctuation model of RNA processing in human mitochondria. Nature 1981, 290: 470–474.
25 Berthier F, Renaud M, Alziari S and Durand R. RNA mapping on Drosophila mitochondrial DNA: precursors and template strands. Nucleic Acids Res 1986, 14: 4519–4533.
26 Zhang DX and Hewitt GM. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. Biochem Syst Ecol 1997, 25: 99–120.
27 Clary DO and Wolstenholme DR. *Drosophila* mitochondrial DNA: conserved sequences in the A + T-rich region and supporting evidence for a secondary structure model of the small ribosomal RNA. J Mol Evol 1987, 25: 116–125.
28 Schultheis AS, Weigt LA and Hendricks AC. Arrangement and structural conservation of the mitochondrial control region of two species of Plecoptera: utility of tandem repeat-containing regions in studies of population genetics and evolutionary history. Insect Mol Biol 2002, 11: 605–610.
29 Kim I, Lee EM, Seol KY, Yun EY, Lee YB, Hwang JS and Jin BR. The mitochondrial genome of the Korean hairstreak, *Coreana raphaelis* (Lepidoptera: Lycaenidae). Insect Mol Biol 2006, 15: 217–225.