Tandem duplication of two tRNA genes in the mitochondrial genome of *Tagiades vajuna* (Lepidoptera: Hesperiidae)

**FANG-FANG LIU**1,*, **YI-PING LI**1,2,*, **IVAN JAKOVLJIĆ**3 and **XIANG-QUN YUAN**1,**

1 Key Laboratory of Plant Protection Resources and Pest Management, Ministry of Education; College of Plant Protection, Northwest A&F University, Yangling, Shaanxi 712100, China; e-mails: hndlf@163.com (F.-F.L.), lyp2003@126.com (Y.-P.L), yuaxq@nwsuaf.edu.cn (X.-Q.Y.)

2 Key Laboratory of Applied Entomology, Northwest A&F University, Yangling, Shaanxi 712100, China

3 Department of Fisheries, Beekeeping, Game Management and Special Zoology, Faculty of Agriculture, University of Zagreb, Svetosimunska cesta 25, 10000 Zagreb, Croatia; e-mail: ivanjakovlic@yahoo.com

**Key words.** Lepidoptera, Hesperiidae, Pyrginae, Eudaminae, *Tagiades vajuna*, mitochondrial phylogenomics, tRNA duplication, butterflies, skippers

**Abstract.** To explore the debated phylogenetic relationship of two Hesperiidae subfamilies, Pyrginae and Eudaminae, and contribute to the understanding of the evolution of mitogenomic architecture in butterflies, we sequenced the complete mitogenome of *Tagiades vajuna*. The mitogenome is a typical circular duplex molecule of 15,359 bp. Apart from the standard 22 tRNAs, it has a tandem duplication of *trnS1*(AGN) and *trnE*, which is unique in lepidopteran insects. Comparison with *Ctenoplistum vasava* indicates that the *trnS1* duplication is not an ancestral state shared with other species of Tagiadiini. Independent origin of the *trnS1* duplications was further confirmed by the reconstruction of the ancestral character state based on the topology of the phylogram. Furthermore, comparative analysis of mitogenomes with and without tRNA duplications indicates that tRNA duplication does not alter the codon usage pattern. The mitogenome has negative AT- and GC-skews, and it is highly A+T-biased (79.7%). The AT-rich (or control) region (283 bp) contains “ATAGA” and “ATTAA” motifs. Regarding the phylogenetic analysis, we found that removal of the third codon position (3CP) from datasets used for the mitochondrial phylogenomics of Hesperiidae is likely to produce results that are more consistent: Pyrginae were rendered paraphyletic by Eudaminae in both analyses of the dataset from which the 3CP was removed (13 PCGs + all RNAs), but inclusion of the 3CP resulted in a destabilized topology, resulting in both monophyly and polyphyley. We conclude that even shallow-phylogenies of insects should pay close attention to compositional and mutational biases in mitogenomes.

**INTRODUCTION**

Skippers (Hesperiidae) are a species-rich family (approximately 3,587 species), which makes up one fifth of the butterfly species in the world. Previous studies have found that neither three mitochondrial genes (Yuan et al., 2015), nor ten nuclear and mitochondrial markers (Sahoo et al., 2016), were capable of providing sufficient phylogenetic resolution to clarify the evolutionary relationships among the Hesperiidae. Particularly difficult to resolve are the relationships among the subfamilies Pyrginae, Eudaminae and Euchromiinae. This led the above authors to propose that future attempts would need much larger datasets.

Due to several advantageous characteristics of mitochondrial genomes, which include small size, abundance in tissues, strict orthology of encoded genes, presence of genes/regions evolving at different rates, uniparental inheritance and general absence of recombination, they are often the tool of choice for resolving phylogenetic relationships among insects (Brown & Wilson, 1979; Ballard & Whitlock, 2004; Gissi et al., 2008; Nelson et al., 2012; Simon & Hadrys, 2013; Wan et al., 2013; Cameron, 2014; Dai et al., 2016). There are some limitations to the application of mitochondrial phylogenomics in insects, such as compositional heterogeneity (Cameron, 2014) and mutation saturation in the third codon position (3CP) of protein-coding genes (PCGs) (Cameron et al., 2012). However, these limitations are relatively well-understood in deep-phylogenies of Insecta lineages (Cameron et al., 2007; Sheffield et al., 2009; Cameron, 2014; Song et al., 2016), the extent of their effects remains incompletely investigated (and occasionally even ignored, see below) in shallow-phylogenies of insects. This might be a consequence of an insufficient

* Contributed equally.

** Corresponding author; e-mail: yuaxq@nwsuaf.edu.cn
number of mitogenomes being available for such phylogenies. The latest two studies of the family Hesperioidea, employing the mitochondrial phylogenomics approach (Zhang et al., 2017b, c), found that the subfamilies Pyrginae and Eudaminae exhibit unresolved polytomies. As we suspect that their results may have been influenced by the aforementioned limitations of this approach, we adopt the most commonly used strategy to alleviate compositional heterogeneity: removal of the 3CP of PCGs. In addition, to improve taxon sampling and phylogenetic resolution, we sequenced the mitochondrial genome of a species belonging to the subfamily Pyrginae, Tagiades vajuna. To assess the effects of this strategy, we also conducted analyses on a dataset comprising all three codon positions of all PCGs. Although the removal of RNAs (rRNAs and tRNAs) has been a common approach used in studies utilizing mitogenomic data to investigate the phylogenetic relationships among the major lineages of Lepidoptera (Kim et al., 2011; Yang et al., 2015), there is no evidence of significant incongruence in the phylogenetic signals between RNAs and PCGs in insects (Cameron et al., 2007; Cameron, 2014). RNAs can actually carry a considerable phylogenetic signal and their inclusion has had positive effects on nodal support for some lineages within Lepidoptera (Wan et al., 2013). Thus we have adopted a strategy of maximizing the amount of data by including all 37 mitogenomic genes: 13 PCGs (excluded/included 3CP), two rRNAs and 22 tRNAs. Beyond mitochondrial phylogenomics, mitogenomic gene duplication and the existence of pseudogenes is of interest in studies of the evolutionary history and mechanisms of gene rearrangement and recruitment (Ye et al., 2016). Although tRNAs are considered to be the most “expendable” among the genes encoded by the mitochondrial genomes and their content and order in mitogenomes is rather variable (Gissi et al., 2008), only four tRNA duplications and/or tRNA pseudogenes are recorded in the superfamily Papilionoidea. Among these four, one species belongs to Lycaenidae, Coreana raphaelis (trnS1 duplication) (Kim et al., 2006), one to Nymphalidae, Acraea issoria (trnl pseudogene) (Hu et al., 2010) and the remaining two are hesperiids: Ochlodes venata (trnL2 pseudogene) and Ctenoptilum vasava (trnS1 duplication and trnL2 pseudogene) (Hao et al., 2012). That for C. vasava, is one of the only two available mitogenomes for the tribe Tagiadinidae. The other one for Ditamio tethys is a standard mitogenome. Therefore, it remains unclear whether this tRNA duplication is an autapomorphy for C. vasava, or whether it is shared by other, yet unsequenced, mitochondrial genomes in the tribe Tagiadinini. To further explore this phenomenon, we have sequenced and characterized the mitogenome of T. vajuna, another Tagiadinini butterfly, predominantly distributed in South and East Asia. Herein, we record a tandem tRNA duplication event, unique among Lepidoptera. We discuss the evolutionary mechanism for the duplication, and explore the correlation between tRNA duplication and codon usage. We also further compare the tRNA duplication events in the mitogenomes of T. vajuna and C. vasava, and investigate tRNA gene duplication in the broad phylogenetic context of the family Hesperioidea.

** MATERIALS AND METHODS **

** Sample collection and DNA extraction **

A single specimen of an adult T. vajuna (Maruyama, 1991) was captured by Shi-Hong Jiang on 14th November 2015 in Yuan-shan Park, Shenzhen City, Guangdong Province, China (22°43′– 23°15′N; 114°14′–115°16′E). The specimen was taxonomically identified by its morphological characteristics (particularly the genitalia) (Yuan et al., 2015), and coxl barcoding (Fig. S1) using 491 of the 493 sequences available for Pyrginae in the BOLD database (Ratnasingham & Hebert, 2013) (two incomplete Pyrginae barcodes were excluded). The complete specimen was immediately preserved in 100% ethanol and stored at −20°C in the Entomological Museum of the Northwest A&F University, Yangling, Shaanxi Province, China. The total DNA was extracted from the thoracic muscles following the manufacturer’s instructions (EasyPure™ Genomic DNA Kit, TRAN, TransGen, Beijing, China). As T. vajuna is an unprotected invertebrate species, no permits were required for this study.

** Sequence analysis **

The complete mitochondrial genome of T. vajuna was sequenced on an Illumina HiSeq2000 system by the Genesky Biotechnologies Inc. company (Shanghai, China). Illumina sequencing data were quality-trimmed with Trimomatic v0.35 (Bolger et al., 2014) and used for mitochondrial DNA genome assembly via a two-step MIRA4/Mitobim combined pipeline (Hahn et al., 2013), which implements a hybrid mapping and assembly approach for the targeted assembly of homologous sequences. The annotation of the mitochondrial DNA sequence was carried out in Geneious 8.1.3 (Biomatters, Auckland, New Zealand), using the mitogenome of another species of Hesperioidea, Parnara gutata (Hesperioidea: Hesperiidae; GenBank: NC_029136) (Shao et al., 2015), as a reference. Protein-coding genes (PCGs) were determined by finding the ORFs (employing codon table 5), and tRNAs (12S and 16S) were identified using the MITOS Web Server (Bernt et al., 2013). Transfer RNAs, including the two duplicated tRNAs, trnL and trnS(AAG), were also identified by MITOS and by manually inspecting the potential cloverleaf secondary structures and anticodons. Finally, all genes were visually inspected against the reference mitogenome via alignments in Geneious. Nucleotide composition and codon usage were calculated using MEGA 6.0 (Tamura et al., 2013). Comparative analysis of the codon usage was carried out on 87 complete lepidoper-teran mitogenomes available from GenBank: all 27 Hesperioidea species randomly chosen from the pool (2–5 species per family), including nine superfamilies and 20 families (Table S2). The mito- genome sequence is deposited in GenBank under the accession number KX865091.

** Phylogenetic analysis **

Phylogenetic analyses were conducted on the newly sequenced T. vajuna mitogenome and 26 complete Hesperioidea mitogenomes retrieved from GenBank. Two species, Ereuna hecate (NC_022685, Pieridae) and Papilio machaon (NC_018047, Papilionidae), were used as outgroups, adding up to 29 mitogenomes in total (Table S1).

Two GUI-based molecular biology tools, MitoTool (Zhang, 2016b) and BioSuite (Zhang, 2016a), developed by our colleague from the Chinese Academy of Sciences, Dong Zhang, were used to manage sequences and generate statistical tables as described before (Li et al., 2017; Zhang et al., 2017a). Fasta files with
nucleotide sequences for all 37 genes (13 PCGs, 2 rRNAs and 22 tRNAs) were extracted from GenBank files using MitoTool. PCGs were aligned in batches with MAFFT integrated into BioSuite, using codon-alignment mode. All RNAs were aligned with Q-INS-i algorithm (which takes secondary structure information into account) incorporated into MAFFT-with-extensions software (Katoh & Standley, 2013). Phylogenetic analyses were conducted using two different datasets: the complete 13 PCGs + all RNAs (named PCGRT dataset), and 13 PCGs with 3CP removed + all RNAs (named PCGI2RT).

Best partitioning strategies and models for the two datasets were selected using PartitionFinder v1.1.1 (Lanfear et al., 2012). We created 16 pre-defined partitions of the two datasets: 13 PCGs + 2 rRNAs + all concatenated tRNAs as a single partition. We utilized the “greedy” algorithm (with branch lengths estimated as “unlinked”) and Bayesian information criterion (BIC) to search for the best-fitting scheme (Table S3). Phylogenetic analyses were performed employing the best-fitting partitioning schemes, using maximum likelihood (ML) and Bayesian inference (BI). The ML analyses were performed using RaxML GUI (Silvestro & Michalak, 2012; Stamatakis, 2014), with an ML + rapid bootstrap (BS) algorithm with 1000 replicates. The BI analyses were performed using MrBayes 3.2.6 (Ronquist et al., 2012) with default settings and 6 × 10^6 MCMC generations (average standard deviation of split frequencies < 0.01, estimated sample size > 200, and potential scale reduction factor ≈ 1). Based on the resultant phylogram, we conducted an ancestral character state reconstruction for the tRNA duplications within the tribe Tagiadiini using the MLGO web server (Hu et al., 2014).

RESULTS AND DISCUSSION

Genome features and characteristics

The complete mitogenome of T. vajuna (Fig. 1, Table 1) is 15,359 bp-long. It contains the standard 13 PCGs, two ribosomal RNAs (12S and 16S) and the non-coding AT-rich region (also called the control region). Intriguingly, two duplicated tRNA genes (trnS1 and trnE) were found in this mitogenome with the help of the MITOS algorithm (discussed in the “tRNA genes” section), thus adding a further 24 tRNA genes. Fourteen genes are transcribed from the N strand and the remaining 25 genes from the J strand. Apart from the two duplicated trn genes, the order of genes in the T. vajuna mitogenome is relatively typical of Lepidoptera (Kim et al., 2009).

All PCGs, including the COI, which usually uses non-standard start codons in this group of animals (Ramírez-Ríos et al., 2016), use the standard ATN and TAA (or its
The very high A+T content of 79.7% is comparable with that recorded for other lepidopteran mitogenomes (Hao et al., 2012; Kim et al., 2014), corresponding well with the particularly high AT content at the first codon position (72.8%) than at the second position (69.9%), but by far the greatest at the third codon position – 91.4% (Table 2). This particularly high background mutational pressure towards A/T nucleotides at the third codon position is common in butterfly mitogenomes (92% on average) (Min et al., 2014).

To further explore this high bias towards A and T nucleotides, and demonstrate the frequency of synonymous codon usage, we have calculated the relative synonymous codon usage (RSCU) values (Fig. 2). UUA (Leu2), UCU (Ser2), CUA (Arg) and CUC (Pro) were the most frequently used codons and CUG, ACG, UGC and AGC the less frequently used. Three families (Leu2, Ile, Phe) account for 34.59% of all codons. Codons ending in A or T were predominant, adding up to 3,390, and accounting for 90.96%. The strong preference for A+T-rich codons over synonymous codons with a lower A+T content in almost all amino acids is observable in Fig. 2. The preference is particularly obvious with a lower A+T content in almost all amino acids is observable in Fig. 2. The preference is particularly obvious with the exception of Pro (Fig. 2), where the AAU codon was used in 95.44% cases, as opposed to only 4.66% for UUG. This prevalence of NNU and RNAA codons, also recorded in other skipper mitogenomes (Hao et al., 2012; Kim et al., 2014), corresponds well with the particularly high AT content at the third codon position.

tRNA genes
The mitogenome of T. vajuna harbours 24 tRNA genes interspersed between rRNAs and PCGs and ranging in length from 61 to 72 bp (Fig. 3, Table 1). Among them, 16 are encoded on the J strand and eight on the N strand. With the exception of trnS1(AGN), which lacks the DHU arm, all of the tRNAs could be folded into cloverleaf secondary structures using MITOS (Fig. 3). The missing DHU stem of trnS(AGN) is an ancestral state in butterflies, including skippers, and probably evolved very early in Metazoa (Garey & Wolstenholme, 1989).

Comparative analysis of the selected 87 lepidopteran mitogenomes indicates that the tandem trnS1-trnE duplication might be unique among the lepidopteran mitogenomes sequenced. Since it is absent from the phylogenetically closely related D. tethys (Fig. 6), it is likely to be specific to the Tagiades genus. The trnS1 duplication, however, was also recorded in two other species: Cnephasia vasava (Hesperidae: Pyrginae) and Coreana raphaelis (Lycaenidae: Theclinae). Although this indicates that it is possible that the trnS1 duplication (-S1_S1_E-) might be a shared ancestral state for the Tagiades species, the gene order in T. vajuna (-S1_E-S1_E-) indicates that we can reject this hypothesis. This was further reinforced by the results of the ancestral character state reconstruction, which indicate the
typical lepidopteran arrangement -S1-E- as the ancestral state for the two Tagiadiini nodes: A1 and A2 (Fig. 6). This confirms the independent origin of the duplicated trnS1 in these two species: -S1-E- → -S1a-S1b-E- in C. vasava, whereas in T. vajuna it results from a duplication of the entire segment: -S1-E- → -S1a-Ea-S1b-Eb- (Fig. 5).

Comparison of the copies of the two duplicated tRNAs within and between the species (Fig. 4) lends further sup-

**Fig. 3.** Predicted secondary cloverleaf structure for the tRNAs of *T. vajuna* mitogenome. Dashes indicate Watson-Crick base pairing; additional sign (+) indicates unmatched base pairing.
port to this scenario: the two trnS1 copies in T. vajuna are identical to each other, but different from the two C. vasava orthologs (77.05 and 88.71% similarity, Fig. 4). Tagiades vajuna trnEa and trnEb copies are not identical, but the differences are relatively minor (91.04% similarity). Comparison with the C. vasava trnE sequences indicates that T. vajuna trnEb is the faster-evolving copy (b = 82.61 and a = 84.85% similarity), which is likely to be a consequence of relaxed mutational constraints afforded by the functional redundancy. However, all four copies can be folded into a functional cloverleaf structure (Fig. 3), which indicates that the duplication is relatively recent and that these tRNAs have probably retained their functionality. The codon usage for serine (AGN) (C. raphaelis, C. vasava and T. vajuna) and glutamic acid (only T. vajuna) encoded with two tRNA copies is analogous among all lepidopterans (Table S2), which indicates that tRNA duplication did not alter the codon usage pattern.

**Intergenic spacers, overlapping sequences and the AT-rich region**

Intergenic spacers in the T. vajuna mitogenome, distributed in 16 regions, with sizes ranging from 1 to 66 bp, add up to 151 bp. Two of these, Spacer 1 (66 bp, located between trnQ and ND2) and Spacer 6 (17 bp, trnS2 and ND1; Table 1), are believed to be a constitutive synapomorphic feature of lepidopteran mitogenomes, and likely to have a functional role (Cameron & Whiting, 2008; Kim et al., 2009, 2010; Sheffield et al., 2008; Taanman, 1999).

There are eleven gene overlaps in the mitogenome, 1 to 8 bp in size, adding up to 36 bp. The two longest overlaps are between trnW / trnC and ND4L / trnT genes (Table 1). Overlapping genes might be a reflection of the selection for a short and economic mitogenome, and they usually involve trn genes, because their sequences are constrained by fewer mutations (Doublet et al., 2015).

![Fig. 4. Comparison of the duplicated T. vajuna tRNAs, trnS1 and trnE, with selected orthologs. Similarity (%) between the sequences is indicated in the matrix on the left, where Tv and Cv are abbreviated names of the two species.](image)

Similar to some other Lepidoptera (Liao et al., 2010), the AT-rich region (283 bp, A+T = 91.5%) is located between rrnS and trnM in the T. vajuna mitogenome (Table 1, Fig. 1). These regions commonly have an ATTA motif followed by several runs of microsatellite-like A/T sequences in other Lepidoptera (Cameron & Whiting, 2008). They also possess an ATAGA motif close to the 5’-end of the 12S rRNA gene, followed by a poly-T stretch of variable length and a poly-A stretch (which can be interrupted or uninterrupted) immediately upstream of trnM (Kim et al., 2014). These two motifs also occur in the AT-rich region of the T. vajuna mitogenome. The poly-T stretch following ATAGA motif was 19 bp-long, whereas the poly-A stretch was comprised of 14 bp, and interrupted by a T base at position 12. The AT-rich region is believed to be involved in the control of transcription in insects (Zhang et al., 1995).

**Phylogenetic analyses**

As both methods (BI and ML) used in the phylogenetic analyses produced concordant topologies using the PCG12RT dataset, only the BI tree is shown (Fig. 6, all remaining phylogenies are shown in Fig. S2). Using the PCGRRT dataset, however, the topologies produced (Fig. S2) were neither concordant with each other, nor with the PCG12RT topology. Hence, we can conclude that removal of the 3CP from datasets used for the mitochondrial phylogenomics of Hesperiidae is likely to produce results that are more consistent. As expected, T. vajuna clustered with the other two Tagiadini species, D. tethys and C. vasava, in all four phylograms (two datasets × two methods) produced. However, phylogenetic relationships among/within the Pyrginae and Eudaminae subfamilies varied among the four phylograms: Pyrginae were rendered paraphyletic by Eudaminae in both analyses of the PCG12RT dataset, but polyphyletic in the ML analysis and monophyletic in the BI analysis of the PCGRRT dataset.

Eudaminae recognized as a new hesperiid subfamily on the basis of a combination of molecular and morphological data by Warren et al. (2009) was quickly disputed as there is no morphological evidence for the monophyly of either Eudaminae or Pyrginae in the new sense (Simonsen et al., 2012). Sahoo et al. (2016) report that the position of Eudaminae in relation to Pyrginae is very variable depending on the methodological approach, with particular emphasis...
on the partitioning schemes. They conclude that this is a reflection of the insufficient information in their molecular dataset (ten nuclear and mitochondrial markers) and propose that future attempts will have to rely on phylogenomic approaches. The mitochondrial phylogenomic approach used in this study, however, did not resolve the controversial phylogenetic relationship of the subfamilies Pyrginae and Eudaminae. As the PCG12RT dataset produced more consistent results and higher nodal support (Fig. S2), we hypothesize that this dataset may have produced results more closely reflecting the actual relationships between the Pyrginae and Eudaminae.

CONCLUSIONS

In this study, we report a tandem duplication of two tRNA genes, unique among all the lepidopteran mitogenomes characterized. Comparative analysis and results of an ancestral character state reconstruction indicate the trnS1 duplication recorded in two species of Tagiadiini (T. vajuna and C. vasava) is not a synapomorphy. Sequencing of further Tagiadiini mitogenomes will be needed to determine whether the tandem duplication of trnS1-trnE recorded in T. vajuna is autapomorphic just for this species or the entire genus (Tagiadi). Apart from the novel tRNA duplication, the mitogenome of T. vajuna has the standard features of Lepidoptera. Although our analyses indicate that the subfamily Pyrginae is most likely paraphyletic, varying topologies produced by different datasets and methods indicate that mitochondrial phylogenomics may not be able to fully resolve the phylogenetic relationships of the subfamilies Eudaminae and Pyrginae. The unstable topologies and weak nodal support recorded in both analyses (BI and ML) of the PCGRT dataset indicate that even shallow-phylogenies of insects should pay close attention to compositional and mutational biases in mitogenomes.

ACKNOWLEDGEMENTS. We thank De-Long Guan and Fan Song for their help with some of the programs used in the analyses. Apart from contributing the programs Dong Zhang also helped us to devise this methodology by contributing valuable ideas. We also thank J.R. Schrock for reviewing the manuscript, as well as three anonymous reviewers, whose constructive comments helped us improve this manuscript. This study was supported by the National Natural Science Foundation of China (Grants No.31272345, 31071693 and 31772503), Chinese Universities Scientific Fund (Grants No. Z1090021710) and Biodiversity Conservation Program of the Ministry of Environmental Protection, China (Grant No. SDZXWJZ01048-2017).

REFERENCES

CAMERON S.L. & WHITING M.F. 2008: The complete mitochondrial genome of the tobacco hornworm, Manduca sexta (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. — Gene 408: 112–123.


Received June 21, 2017; revised and accepted September 25, 2017 Published online October 12, 2017

Supplementary files:
Table S1 (http://www.eje.cz/2017/052/S01.xlsx). List of the species included in this study.
Table S2 (http://www.eje.cz/2017/052/S02.xlsx). Codon usage for glutamic acid and serine (AGN) among 87 selected Lepidoptera.
Table S3 (http://www.eje.cz/2017/052/S03.docx). The best partitioning scheme selected by PartitionFinder for the two datasets. Fig. S1 (http://www.eje.cz/2017/052/S04.pdf). Phylogenetic trees based on 491 cox1 barcoding sequences.
Fig. S2 (http://www.eje.cz/2017/052/S05.pdf). Phylogenetic trees based on all four analyses (2 datasets × 2 analytical methods).