Gene 576 (2016) 404–411

Contents lists available at ScienceDirect

Gene

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

Research paper

The complete mitogenomes of *Calameuta filiformis* (Eversmann, 1847) and *Calameuta idolon* (Rossi, 1794) (Hymenoptera: Cephidae): The remarkable features of the elongated A + T rich region in Cephini

E. Mahir Korkmaz [a,*], Mahir Budak [a], Merve Nur Ördek [a], Hasan Hüseyin Başıbüyük [b]

[a] Department of Molecular Biology and Genetics, Faculty of Science, Cumhuriyet University, Sivas, 58140, Turkey
[b] Department of Biology, Faculty of Science, Cumhuriyet University, Sivas, 58140, Turkey

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 17 October 2015
Accepted 20 October 2015
Available online 26 October 2015

Keywords:
Mitochondrial genome
Sawflies
Replication fork barrier
Tandem repeats
rRNA-like sequences
tRNA-like sequences
Symphyta

ABSTRACT

Two complete mitogenomes of the stem borers, *Calameuta filiformis* and *Calameuta idolon*, and the complete A + T-rich region of *Trachelus iudaicus* (Hymenoptera: Cephidae), are reported. The mitogenomes of these species are the longest reported from hymenopterans to date. A remarkable increase in length of the A + T-rich region, the longest for Hymenoptera, was found and compared across the tribe Cephini. The presence of the tRNA- and rRNA-like sequences were reported in the A + T-rich region of sawflies and they were suggested to play a role in replication and/or transcription. The long and short tandem repeats were orderly located in both sides of the A + T-rich region producing stable secondary structures. We suggest that the short tandem repeats are likely to function as a replication fork barrier.

1. Introduction

The mitogenome has been one of the main source of information for insect studies for last several decades. At present, the mitogenome of at least one species for each order of insect is represented in the databases (NCBI, June 2015). The mitogenome information is widely utilised in understanding of insect taxonomy, and their evolutionary history and life strategies (Cameron, 2014). A typical insect mitogenome is a double strand circular molecule and ranges between 14 and 20 kb in length. It is generally fixed to 37 genes containing 13 protein coding genes (PCGs), two rRNAs and 22 tRNAs (Boore, 1999). In addition to its gene content, mitogenomes contain a major non-coding segment (also known as A + T-rich region or control region) responsible for initiation and regulation of replication and/or transcription (Saito et al., 2005; Wolstenholme, 1992; Zhang and Hewitt, 1997).

The location of the replication origin and the structural and/or regulatory motifs of the A + T-rich region has been well described in insect mitogenomes (Saito et al., 2005; Zhang and Hewitt, 1997; Zhang et al., 1995). Despite presence of more information on replication, relatively little is known about the transcription features of insect mitogenomes (Beckenbach and Stewart, 2009; Wang et al., 2013). Transcription of insect mitogenomes starts from two promotors on the majority strand and from one promotors on the minority strand, and mRNAs are mostly transcribed from the gene blocks located in the majority and minority strands, respectively (Torres et al., 2009).

Hymenoptera is one of the large insect order including 150,000 species with very diverse life strategies and considered as the early diverged lineage of holometabolous insects (Sharkey, 2007 and reference therein). So far, complete or nearly-complete mitogenomes of 81 hymenopteran species have been reported (NCBI, June 2015). Of these, only three complete and four nearly-complete mitogenomes are from the suborder Symphyta, the sawflies. The mitogenome size ranges from 15,418 to 19,339 bp in sawflies and length variation is mainly due to variation in the copy numbers of tandem repeat sequences (TDRs) (Wei et al., 2010a, 2010b; Zhang et al., 1995). The longest A + T-rich region (3872 bp, NC012688) known from the Hymenoptera belongs to a stem borer species *Cephus cinctus* Norton, (sequenced by Dowton et al., 2009), an economically important pest in North America with

Abbreviations: *Cl. idolon*, *Calamauta idolon*; *Cl. filiformis*, *Calamauta filiformis*; *T. iudaicus*, *Trachelus iudaicus*; *C. cinctus*, *Cephus cinctus*; bp, base pair; PCGs, protein coding genes; rRNA, ribosomal RNA; tRNA, transfer RNA; A + T-rich region, adenine + thymine-rich region; TDRs, tandem repeat sequences; PCR, polymerase chain reaction; NGS, next generation sequencing; ORF, open reading frame; IQM cluster, *trnI*, *trnQ*, *trnM* gene cluster; *atp6* and *atp8* genes, ATP synthase subunits 6 and 8; *cox1–3* genes, subunits 1–3 of cytochrome c oxidase; *nad1–6* and *nad4l*, NADH dehydrogenase subunits 1–6 and 4l; *rrnS* and *rrnL* genes, small and large subunits of ribosomal RNA; AT, adenine and thymine; GC, guanine and cytosine; DHU, dihidroxyuridine; TΨC arm, pseudouridine arm; H strand, majority strand; L strand, minority strand.

* Corresponding author.
E-mail address: ekorkmaz@cumhuriyet.edu.tr (E.M. Korkmaz).

http://dx.doi.org/10.1016/j.gene.2015.10.050
0378-1119/

phytophagous life strategy (Ivie, 2001). Despite a frequently observed trend of size reduction in insect mitogenomes, the significance of increased A + T-rich region size in Symphyta is unknown and requires a detailed investigation (Boore, 1999; Burger et al., 2003).

Here, we report two complete mitogenomes from the stem borers, *Calameuta filiformis* (Eversmann, 1847) and *Calameuta idolon* (Rossi, 1794). We also report the A + T-rich region of a third cephini species *Trachelus iudaicus* (Konow, 1907) and compare A + T rich region evolution across the Cephini and to other sawflies. We describe the location of the replication origin and the structural and/or regulatory motifs of the A + T-rich region in Cephini for the first time. We have also identified tRNA-and rRNA-like sequences and discussed their likely origins and potential functions in replication and transcription. We suggest that the TDRs located to both sides of "the control region" may function as a replication fork barrier and as binding sites for proteins during replication and/or transcription.

2. Material and methods

2.1. Specimens and DNA isolation

Specimens of *Cl. filiformis*, *Cl. idolon* and *T. iudaicus* were taken from ethanol preserved material of the Entomological Collection of Cumhuriyet University, Sivas (ECCUS). For each species, whole genomic DNA was extracted from hind leg of a specimen by a simple salting out procedure (Sunnucks and Hales, 1996).

2.2. Mitogenome amplification and sequencing

The mitogenomes of *Calameuta* species were amplified using 15 standard and one long primer pairs previously reported (Korkmaz et al., 2015) (see also Sup. Table 1). Both standard and long PCRs were performed using a BioRad T-100 thermal cycler (Korkmaz et al., 2015). All PCR products were purified by polyethylene glycol (PEG) precipitation (Paithankar and Prasad, 1991). Purified standard PCR amplicons were Sanger-sequenced in both directions using the same primers as in PCR reactions at Macrogen Inc. After failure to sequence the long PCR products using traditional methods, the purified products were sequenced using the Illumina HiSeq 2000 next generation sequencing (NGS) platform using 250 bp paired end reads, conducted at DONE Genetics Inc., Turkey.

2.3. Mitogenome assembly, annotation and analyses

Sequence chromatograms for the standard PCR amplicons were individually aligned and compiled into contigs using the CodonCode Aligner v4.1.1 (CodonCode Corporation). NGS reads were imported into Geneious R8 (http://www.geneious.com) (Kearse et al., 2012). For the initial quality control reads, raw sequence reads ranging between 35 bp and 251 bp in length were trimmed from both the 3′ and 5′ ends. Any reads less than 100 bp were also filtered out to remove low quality scores. Post quality control reads were ~242 bp in length with 16% standard deviation. High quality reads were then assembled into contigs using three different approaches for further mapping the entire A + T-rich region: (i) a reference assembly using the mitogenome of *C. cinctus* (NC012688) under the parameters of 'medium–low sensitivity' and 'iterate up to 5 times', (ii) *de novo* assembly using the parameter of 'highest sensitivity' and then the obtained contigs were mapped with the reference mitogenome generated by the first approach, and (iii) an assembly produced using the *rrnS* and *trnQ* sequences located on both sides of the A + T-rich region (obtained from standard sequencing for each species), under the parameters of 'medium–low sensitivity' and 'iterate up to 5 times'. The sets of selected assemblies produced by these approaches were aligned, manually compared and finally compiled into a single contig. Repetitive sequence located at both the 5′ and 3′ ends of the A + T rich region hampered the generation of sequences covering the entire A + T-rich region. Therefore, this partial consensus sequence was scanned using Tandem Repeat Finder v4.07b (Benson, 1999) to identify repeat sequence motifs. To estimate the number of repeat units, long PCR amplicon sizes were taken into account and the identified long (at the 5′ end) and short repeated sequences (at the 3′ end) were iteratively scanned and manually examined in the sets of selected assemblies. Finally, contigs from both Sanger and NGS were integrated by software-mediated recognition of the matched sequences between contigs using BioLign v4.0.6 (http://en.bio-soft.net/dna/BioLign.html). The complete mitogenome sequences of the two *Calameuta* species and the A + T-rich region of *T. iudaicus* have been deposited in the GenBank under the accession numbers KT260168 (*Cl. idolon*), KT260167 (*Cl. filiformis*) and KT260169 (*T. iudaicus*).

tRNA genes were found by their putative secondary structure and anticodon sequence using tRNAscan-SE server (Lowe and Eddy, 1997) and ARWEN v1.2 (Laslett and Canbäck, 2008) with the mito/chloroplast genetic code and the default search options. PCGs were analysed by ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) using the invertebrate mitochondrial code. In case of failure in the ORF Finder, they were aligned and checked manually on the basis of the homologous gene sequences from other known stem borers (NC012688, KM377623 and KM377624) using MEGA v6 (Tamura et al., 2013) and BLASTP searches against available insect mitogenomes. The boundaries of the rRNA genes were inferred from boundaries of adjacent genes (Dowton et al., 2009). Their secondary structures were constructed with reference to other reported stem borer species (Korkmaz et al., 2015) and XRNA v1.1.12b (http://rna.ucsc.edu/rnacenter/xrna/xrna.html) was used to draw the folding structure based on the results of the CRW site (Cannone et al., 2002). Basic nucleotide composition statistics were also calculated by MEGA v6. Strand asymmetry was calculated using the formulas AT-skew = [A − T] / [A + T] and GC-skew = [G − C] / [G + C] (Perna and Kocher, 1995) to measure the base compositional difference between the different strands, the genes coded on the alternative strands and degenerated codon positions.

2.4. A + T-rich regions

In addition to the A + T-rich region of Cephini (*Calameuta*, *Trachelus* and *Cephus*), two other sawflies with complete A + T-rich region (*Allantus luctifer*: Tenthredinidae and *Orussus occidentalis*: Orussidae) were also included in the analyses of this region. The boundaries of the A + T-rich region in each species were determined by considering the adjacent genes. tRNA-like structures were predicted using ARWEN with default options. The 'invertebrate mitochondrial' genetic code was taken as reference to find ORF-like sequences under the ORF Finder. The rRNA-like sequences were searched based on homology using a "selfblast" parameter by not filtering low complexity regions in NCBI database. To infer whether the predicted tRNA residues are functional or not, the conserved secondary structures and positions of the tRNA were evaluated, their COVE values calculated and their similarities to tRNAs investigated. The identified ORF-like sequences were translated into amino acid sequences and subsequently a homology search was applied under BLASTP search in NCBI database by not filtering low complexity regions from nucleotide sequences. The number, position and length of the repeated sequences were searched by Tandem Repeat Finder v4.07b using default setting under basic option. To see whether the tandem repeats include conserved nucleotides and secondary structures or not in genera and tribe level, the sequences were aligned in MEGA v6.0. Uncorrected p-distance values were calculated for the sequence pairs within species and among species. The putative secondary structures were predicted under Quikfold interface of the DINAMelt Server using the following parameters: linear sequence, DNA folding form with a folding temperature of 25 °C (Markham and Zuker, 2005; Zuker, 2003). Finally, the remaining sequence of A + T-rich region was scanned by eye in terms of likely replication origin(s) and conserved structural and/or regulatory elements (Zhang and Hewitt, 1997).