Heliyon

Received:
25 June 2018
Revised:
6 September 2018
Accepted:
8 November 2018

Cite as: Sarp Kaya, Battal Çıplak. Possibility of numt co-amplification from gigantic genome of Orthoptera: testing efficiency of standard PCR protocol in producing orthologous COI sequences.
Heliyon 4 (2018) e00929.
doi: 10.1016/j.heliyon.2018.e00929

Possibility of numt co-amplification from gigantic genome of Orthoptera: testing efficiency of standard PCR protocol in producing orthologous COI sequences

Sarp Kaya [a], **Battal Çıplak** [b,*]

[a] *Mehmet Akif Ersoy Üniversitesi, Burdur Vocational School of Healt Services, Burdur, Turkey*

[b] *Department of Biology, Faculty of Science, Akdeniz University, 07058, Antalya, Turkey*

[*] Corresponding author.

E-mail address: ciplak@akdeniz.edu.tr (B. Çıplak).

Abstract

Mitochondrial DNA has been the preferential genome biodiversity studies. However, several factors contribute to its inadequacy. Numts constitute one of the main complications that prevent obtaining orthologous mitochondrial sequences. Orthoptera have been a model group in numt studies because of their huge genome size. In this study we aimed to; (i) test efficiency of standard PCR protocol in producing orthologous sequences of cytochrome C oxidase, (ii) study presence/absence of numts in several unstudied Orthoptera species, (iii) test if there is a threshold between the length of mtDNA targeted for amplification and possibility of encountering numts, and (iv) estimate reliability of the sequences in databases in light of these findings. For these aims we studied 38 species of Orthoptera representing different sublineages and genome sizes. DNA extracted from each sample was used to amplify five different fragments of COI region by standard PCR protocol. Sequenced PCR amplicons were checked for numt possibility by several different numt criteria. No sequences without numt signs

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

were obtained for the first fragment. The number of samples with numt signs for the other four fragments differed between the suborders Ensifera and Caelifera. The percentage of samples with numt signs was higher in Caelifera than Ensifera for all fragments. The numt percentage considerably decreased for the longest two fragments. Numts are more prevalent in families with larger genome size. We arrived at the following conclusions: (i) numts are common in all members of Orthoptera, but, their prevalence differs among intra-lineages, especially more prevalent in Caelifera, (ii) there seems a correlation between numt rate and genome size, (iii) there is no threshold to avoid numt co-amplification, but, a 1,000 bp length may be a threshold for Ensifera, (iv) Folmer region of COI doesn't seem an appropriate marker for animal barcoding. Additionally, a phylogenetic tree produced from the numt sequences of fragment four detected in genus *Anterastes* suggested a paleonumt gained in generic ancestor a 3.5−4 times slower divergence rate for numt sequences.

Keywords: Molecular biology, Genetics, Evolution, Zoology, Systematics, Ecology

1. Introduction

The past three−four decades witnessed a revolution in population biology, evolution, phylogeny, and taxonomy by using DNA sequence data. Mitochondrial DNA (mtDNA) has been the preferred genome because of several advantages that make it well suited for molecular studies (Avise, 2000, 2009). It presents in a large number of copies yielding significant amounts of genomic DNA, its high mutation rate and small effective population size often make it an informative genome regarding evolutionary patterns and processes, and it is suitable for molecular clock estimations due to the nature of its inheritance (Ballard and Rand, 2005; White et al., 2008; Avise, 2000, 2009). Until the recent development of next-generation sequencing, DNA sequencing for PCR products was done by Sanger sequencing. This method limits the length of DNA fragments that can be analysed, and thus only particular regions of mtDNA have traditionally been studied.

The cytochrome C oxidase subunit I (COI) has possibly been the most commonly studied marker. But, the popularity of COI is mainly because of being used as the maker for DNA barcoding of animal diversity (Hebert et al., 2003a, 2003b, 2010). Its suitability for barcoding is still an ongoing debate. The c. 680 bp length of COI has been sufficiently applied to some groups (Lopez et al., 1997; Barrett and Hebert, 2005; Hubert et al., 2008; Chen et al., 2011) while uninformative for some others (Tautz et al., 2003; Moritz and Cicero, 2004; Rubinoff et al., 2006; Song et al., 2008; Buhay, 2009; Cristiano et al., 2012; Leite, 2012). Independent of these debates, the present databases include huge numbers of COI sequences from different lineages and from different parts of the world (e.g. BOLD) and

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

even several taxa, especially those suggested to be cryptic species, have been defined using these data. Thus, the reliability of sequence data of COI presently uploaded to databases requires caution. There are several factors causing inadequacy of mtDNA in general and COI specifically such as male-biased gene flow, selection on any mtDNA nucleotide(s) (as the whole genome is one linkage group), retention of ancestral polymorphism and introgression following hybridization (Moritz and Cicero, 2004). Additionally, multiple copies of mtDNA within a cell (heteroplasmy), bacterial infection biasing mtDNA variation and especially nuclear integration of mtDNA (or numts) may prevent amplification and identification of the orthologous sequences of COI or any other fragment of mtDNA (Song et al., 2008; Gaziev and Shaikhaev, 2010). Among these, possibly presence of the numts is the most prevalent reason preventing to obtain the exact mitochondrial sequence especially in Orthoptera or insects (Song et al., 2008, 2014; Moulton et al., 2010; Antunes and Ramos, 2005; Behura, 2007; Leite, 2012; Pons and Vogler, 2005; Sunnucks and Hales, 1996) as well as among other animals (Cao et al., 2011; Rawlings et al., 2010; Schmitz et al., 2005; Soto-Calderón et al., 2012; Triant and De Woody, 2008; Williams and Knowlton, 2001; Kim et al., 2006).

The order Orthoptera is a model group for numt studies. *Locusta migratoria* was one of the first species in which numts were detected (Gellissen et al., 1983), prior to the use of the term numt by Lopez et al. (1994). Later, numts were detected in several other orthopteran species belonging to nine different genera (Zang and Hewitt, 1996a; Vaughan et al., 1999; Bensasson et al., 2000, 2001a), all of which were from Acrididae. Contemporarily with those studies, some reviews which relied on the numt data from orthopteran studies appeared (Zang and Hewitt, 1996b, 2003; Bensasson et al., 2001b). Studies on Orthoptera have triggered similar research in other taxa, and numts were found to be prevalent in almost all eukaryotic organisms (Richly and Leister, 2004; Hazkani-Covo et al., 2010). Later, numts were used to define characters and to clarify phylogenetic inconsistencies suggested by paralog sequences (Berthier et al., 2011; Song et al., 2013). During the course of this study, two comprehensive studies appeared presenting several detailed aspects of numts in 28 species of Orthoptera (Moulton et al., 2010; Song et al., 2014). Of these 28 species, 19 are from Caelifera and 9 are from Ensifera representing different family of their suborder. Song et al. (2014) reported that numts are prevalent across all lineages of the order and numt accumulation is a continuous process. Thus, each genome contains numts of different ages.

Orthoptera are of special interest with respect to numt studies because of their genomic characteristics. There seems to be a positive correlation between the haploid genome sizes (C-values) and numt amount/prevalence (Bensasson et al., 2001b; Hazkani-Covo et al., 2010; Song et al., 2014) in eukaryotes. Orthopteran lineages have giant genomes, with C-values varying from 1.55 (*Hadenoecus subterraneus*, Raphidiophoridae) to 16.93 (*Podisma pedestris,* Acrididae). The smallest

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

genome is ten times larger than genome size of *Drosophila melanogaster* (Bensasson et al., 2001b; Hanrahan and Johnston, 2011; Song et al., 2014; Gregory, 2017). But, the C-values vary considerably within Orthoptera itself. The largest genome is approximately 11 times larger than the smallest one. The genome size of species in Acrididae is especially large with C-values mostly >10, while that of Raphidiophoridae and Gryllidae is around 1.5−2.5 (Bensasson et al., 2001b; Song et al., 2014; Hanrahan and Johnston, 2011; Gregory, 2017). Bensasson et al. (2001b) reported that the rate of DNA loss due to deletion is much slower in the brown mountain grasshopper *Podisma pedestris* compared to *Drosophila* or the cricket *Laupala*, and this is possibly one of the main results of genomic gigantism in Orthoptera. Considering the suggestion that there is a correlation between the genome size and numt abundance/prevalence (Bensasson et al., 2001b; Hazkani-Covo et al., 2010), the expectation would be a considerable variation among different members of Orthoptera in respect to numt content/prevalence. However, Song et al. (2014) reported that there is no difference between lineages of the order in this respect and this statement constitutes a conflict with the assumption that larger genomes include more numts. The most important implication is that numts constitute a massive handicap for producing orthologous sequences from orthopteran taxa.

Detection of numts in different organisms showed that methodologically avoiding numts is challenging. Sorenson and Quinn (1998) were possibly the first to review the methods of avoiding numts. They listed the following measures; (i) using fresh material to avoid genomic degeneration, (ii) studying purified mtDNA to totally eliminate nuclear genome, (iii) studying mitochondria-rich tissues to increase abundance of mt-genome, (iii) using taxon-specific primers to benefit from different evolutionary rate of mt- and nuclear genomes, and (iv) to amplify the entire mtDNA or large portions of it using protocols for extended or long-PCR. Later studies also considered these methods to avoid numt co-amplification (Bensasson et al., 2001b; Zhang and Hewitt, 2003; Triant and De Woody, 2007; Leite, 2012). Optimizing template concentration after template preparation (dilution effect) (Malik et al., 2016) and selective enrichment of the marker under the study (Wolff et al., 2012) were suggested as further methods to reduce numt co-amplification. Finding fresh material may not be always possible and also does not guarantee avoidance of numt co-amplification. Studying purified mitochondria is a time, effort and cost-consuming technique. Studying mitochondria-rich tissue is widely applied at least in Orthoptera as the muscle-rich hind femur is the most frequently used source for DNA isolation, but it does not guarantee no amplification of numts in Orthoptera. Moulton et al. (2010) designed taxon-specific primers for barcoding region of COI still, a considerable amount of numts were encountered. However, Malik et al. (2016) had considerable success avoiding numt amplification by optimizing template concentration and using specific primers for mouse. Amplifying total mt-genome or long PCR for around 4−5 kbp are also promising methods (Triant and De Woody, 2007), but these

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

are expensive and time-consuming methods. On the other hand, a considerable amount of numts are longer than 4 kb or even some are about total mt-genome (Lopez et al., 1996; Dayama et al., 2014). It is widely known that the numt kinds and numbers reduce by the length of targeted sequence increased (Richly and Leister, 2004; Triant and De Woody, 2007; Gaziev and Shaikhaev, 2010). However, no study specifically tests a length threshold that prevents or significantly reduces numt co-amplification, especially for the barcoding marker COI.

There are three general aims of this study. Numts constitute the main threat in obtaining orthologous copies of mitochondrial genes. This is a handicap not only for traditional Sanger sequencing, but also for reads produced by next-generation sequencing. Thus, defining the numt abundance and their nature across lineages still has special importance. Although, it is suggested that there is no difference in numt prevalence between sublineages of Orthoptera (Bensasson et al., 2001a; Song et al., 2014) the genomic size variation within the order still requires testing in unstudied taxa. Additionally, Orthoptera constitute a model group in numt studies and data from Orthoptera has a potential of being generalized to other organisms. Thus, the first aim is to study presence/absence of numts in several unstudied Orthoptera species belonging to different genera, sub-families, and families.

Numts may differ in length and be as large as the full length of mt-genome (e.g. Lopez et al., 1996; Du and Qin, 2015; Sun and Yang, 2016). However, numt sequences of a particular gene are generally shorter than the respective mitochondrial sequence in length, thus, the possibility of numts should decrease with increased length of the targeted mitochondrial marker (Richly and Leister, 2004; Hazkani-Covo et al., 2010; Gaziev and Shaikhaev, 2010). Although, search by softwares in total genome deposited electronic medium provides a considerable knowledge for the length of numts, there is no study specifically testing a length threshold in standard PCR protocols. Thus, the second aim of the present study is to test if there is a threshold between the length of mtDNA targeted to be amplified, and a possibility of encountering numts. More importantly, this threshold may differ between lineages especially in correlation with genome size and Orthoptera constitute the most appropriate group to test this assumption.

Presently there is a huge number of COI sequences uploaded in public databases and most of these have the limited length, generally about the length of barcoding region. It is known that the possibility of numts in the existing data should not be ignored (e.g. Bertheau et al., 2011; Song et al., 2008). However, testing the correlation between the length of targeted marker to be amplified and the possibility of numt co-amplification by standard DNA isolation and PCR procedure will allow estimating the reliability of the sequences in databases. Thus, this experimental was made to address this question.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Materials and methods

2.1. Taxon sampling

We selected a wide range of taxa representing different lineages/sublineages of Orthoptera to achieve the above mentioned aims. In previous numt studies on orthopterans there is a bias toward short-horned grasshoppers of Caelifera and a limited number of Ensifera species (Zang and Hewitt, 1996a,b; Bensasson et al., 2001a; Song et al., 2008, 2014; Moulton et al., 2010). Thus, in the present study taxa sampling is biased toward Ensifera. In total, 38 species were studied (Table 1). Twenty six of these belonged to Ensifera and 12 to Caelifera. The 26 species of Ensifera represent 13 genera classified under five families whereas12 species of

Table 1. List of the Orthoptera taxa examined.

Suborder	Family	Subfamily	Genus	Species
ENSIFERA	Gryllidae	Gryllinae	*Gryllus*	*bimaculatus*
	Gryllotalpidae		*Gryllotalpa*	*gryllotalpa*
	Rhaphidophoridae	Dolichopodinae	*Dolichopoda*	*subordoni*
				lycia
		Troglophylinae	*Troglophylus*	*ferzene*
				gajaci
	Schizodactylidae	Schizodactylinae	*Schizodactylus*	*inexpectatus*
	Tettigoniidae	Bradyporinae	*Callimenus*	*conophalus*
				toros
		Phaneropterinae	*Leptophyses*	*albivittata*
			Tylopsis	*lilifolia*
		Saginae	*Saga*	*ephipigera*
				cappadocia
		Tettigoniinae	*Anterastes*	*uludaghensis*
				ucari
				disparalatus
				antitauricus
				serbicus
				burri
				niger
			Psorodonotus	*caucasicus*
				macedonicus
			Platycleis	*affinis*
				armeniaca
			Anadolua	*davisi*
				burri
CCAELIFERA	Acrididae	Cyrtacanthacridinae	*Schistocerca*	*gregaria*
			Anacridium	*aegyptium*
		Oedipodinae	*Oedalus*	*decorus*
		Gomphocerinae	*Stenobothrus*	*fischeri*
			Chorthippus	*paralallelus*
			Dociostaurus	*marrocanus*
				anatolicus
				brevicolis
	Pamphagidae	Pamphaginae	*Nocaracris*	*cyanipes*
			Paranocaracris	*citripes*
			Acinipe	*davisi*
	Pyrgomorphidae	Pyrgomorphinae	*Pyrgomorpha*	*guentheri*

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Caelifera represent 10 genera and three families (see Table 3 in Result section). None of the 38 species has been studied previously for numt studies. The taxonomic diversity also represents the genome size variation within Orthoptera (Hanrahan and Johnston, 2011; Gregory, 2017).

2.2. DNA extraction and PCR procedure

For amplification, we targeted five different and overlapping fragments of COI using five different couples of primers (Table 2 and Fig. 1; F1, F2, F3, F4 and F5 hereafter). We studied 10 specimens per genus (exceptionally 20 for *Anterastes*) and used the same DNA extract of the same sample for each of five primer couples (for each of F1-F5) to standardize PCR amplification or to minimize the possible errors in laboratory applications. Sequencing both strands of amplicons is another measure to minimize the errors.

Specimens were preserved in 96% ethanol and kept at -20 °C in the Orthoptera collection at Department of Biology, Akdeniz University Antalya-Turkey. Total DNA was extracted from hind femurs with the salt-isopropanol extraction method (Aljanabi and Martinez, 1997). Each amplification was performed in a 50 μl volume containing 0.3 μl of each primer (100 μM), 1 μl dNTP mix (10 mM), 2 μl, 50 mM $MgCl_2$, 5 μl 10X Platinum PCR buffer (containing 200 mMTris−HCl [pH 8.4], 500 mMKCl), 1.25 U Platinum TaqDNA polymerase (Invitrogen), and 50 ng template DNA. Temperature cycling was carried out in a BioRAD Mastercycler. Cycling conditions were 2 min denaturing at 95 °C; (40 s at 95 °C, 30 s at 49 °C, and 50, 1.20 s at 72 °C) x 35 for each fragment separately. PCR products were purified and sequenced in both directions. Double-stranded sequence analysis (performed on a 23 ABI 3730XL DNA analyzer) and purifications and sequencing were made through the intake from the Macrogen sequencing service (Macrogen Inc., Amsterdam, Netherlands).

Table 2. The primer couples used to amplify five fragments mentioned in Fig. 1.

Fragment	Primers	Sequence of the primers	Reference
F1	1460	5′-TACAATCTAACACCTAAATAATTCAGCC-3′	Zang and Hewitt (1996a,b)
	UEAI	5′- GAATAATTCCCATAAATAGATTTACA-3′	
	UEA2	5′- TCAAGATAAAGGAGGATAAACAGTTC-3′	
	UEA2d	5′- GMWARWGGWGGWGGRTAWACWGTTCA-3′	
F2	LCO1490	5′- GGTCAACAAATCATAAAGATATTGG-3′	Folmer et al., (1994)
	HCO2198	5′- TAAACTTCAGGGTGACCAAAAAATCA-3′	
F3	UEA4	5′- AATTTCGGTCAGTTAATAATATAG-3′	Zang and Hewitt (1996a,b)
	2191	5′- CCCTGGTAAAATTAAAATATAAATCTTC-3′	Simon et al. (1994)
F4	2183	5′- CAACATTTATTTTGATTTTTTGG -3′	
	3014	5′- TCCAATGCACTAATCTGCCATATTA -3′	
F5	1718	5′-GGRGGATTTGGAAATTGACTWGTTCC-3′	Simon et al. (1994)
	3014	5′- TCCAATGCACTAATCTGCCATATTA -3′	

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 3. The number and percentage of the samples with and without numt signs per fragments (F2-F5) and per genus, family, and suborder. No samples of F1 was sequenced as all produced multiple electrophoretic bands (N: number of samples examined; N-C: number of chromatograms obtained; N-WNS: number of sequences without numt signs; N-NS: number of sequences with numt signs; * out of total number of specimens examined; ** out of total number of chromatograms).

Taxa			N	Fragment 2			Fragment 3			Fragment 4			Fragment 5		
				N-C	N-WNS	N-NS	N-C	N-WNS	N-NS	N-C	N-WNS	N-NS	N-C	N-WNS	N-NS
Ensifera	Gryllidae	*Gryllus*	10	4	-	4	-	-	-	2	2	-	2	2	-
	Gryllotalpidae	*Gryllotalpa*	10	7	6	1	-	-	-	2	2	-	3	3	-
	Rhaphidophoridae	*Dolichopoda*	10	9	6	3	-	-	-	8	8	-	5	5	-
		Troglophilus	10	9	6	3	-	-	-	5	5	-	6	6	-
	Schizodactylidae	*Schizodactylus*	10	-	-	-	-	-	-	4	4	-	3	2	1
	Tettigoniidae	*Callimenus*	10	7	3	4	2	-	2	10	10	-	7	7	-
		Leptophyes	10	-	-	-	-	-	-	6	6	-	2	2	-
		Tylopsis	10	7	3	4	1	-	1	7	6	1	10	10	-
		Saga	10	10	6	4	1	-	1	3	3	-	10	9	1
		Anterastes	20	14	8	6	-	-	-	12	6	6	18	18	-
		Psorodonotus	10	5	-	5	-	-	-	8	8	-	10	10	-
		Platycleis	10	10	8	2	1	1	-	1	1	-	10	10	-
		Anadolua	10	9	3	6	-	-	-	6	6	-	7	7	-
		TOTAL	**140**	**91**	**49**	**42**	**5**	**1**	**4**	**74**	**67**	**7**	**93**	**91**	**2**
		%		**65***	**53,85****	**46,15****	**3,57***	**20****	**80****	**52,86***	**90,54****	**9,46****	**66,429***	**97,85****	**2,15****

(*continued on next page*)

https://doi.org/10.1016/j.heliyon.2018.e00929

2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 3. (*Continued*)

Taxa			N	Fragment 2			Fragment 3			Fragment 4			Fragment 5		
				N-C	N-WNS	N-NS	N-C	N-WNS	N-NS	N-C	N-WNS	N-NS	N-C	N-WNS	N-NS
Caelifera	Acrididae	*Schistocerca*	5	1	-	1	2	2	-	2	2	-	1	1	-
		Anacridium	5	3	-	3	4	2	2	3	3	-	3	3	-
		Oedalus	10	3	1	2	3	-	3	3	1	2	-	-	-
		Stenobothrus	10	2	-	2	3	-	3	2	1	1	1	1	-
		Chorthippus	10	4	-	4	3	-	3	4	-	4	2	-	2
		Dociostaurus	10	7	2	5	10	2	8	4	3	1	4	2	2
	Pamphagidae	*Nocaracris*	10	8	5	3	8	1	7	5	3	2	6	6	
		Paranocaracris	10	9	-	9	1		1	6	2	4	3	2	1
		Acinipe	10	-	-	-	2	-	2	-	-	-	4	4	-
	Pyrgomorphidae	*Pyrgomorpha*	10	10	-	10	1	-	1	10	3	7	8	2	6
		TOTAL	**90**	**47**	**8**	**39**	**37**	**7**	**30**	**39**	**18**	**21**	**32**	**21**	**11**
		%		**52,22***	**17,02****	**82,98****	**41,11***	**18,92****	**81,08****	**43,33***	**46,15****	**53,85****	**35,56***	**65,63****	**34,38****

https://doi.org/10.1016/j.heliyon.2018.e00929

2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

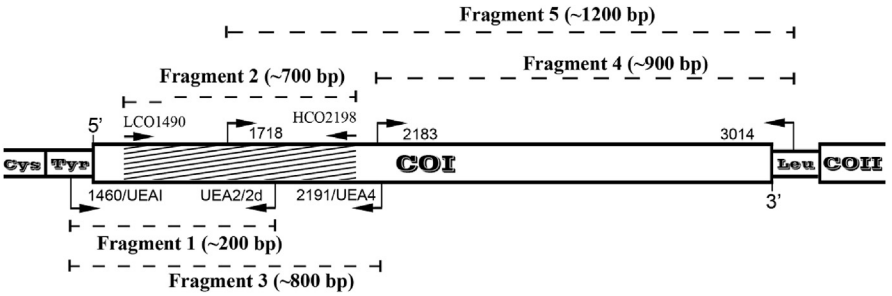

Fig. 1. The amplified 5 different fragments (F1-F5) of COI and the primers used.

2.3. Numt detection

We applied several criteria to detect numts from our PCR products, chromatograms or sequences. First, the PCR amplicons producing a single and prominent electrophoresis bands were sequenced while the others not. Although ghost bands can occur, because of other experimental reasons, numt co-amplification is still a potential reason for dirty PCR products (Zang and Hewitt, 1996b; Kolokotronis et al., 2007; Triant and De Woody, 2007; Moulton et al., 2010). Second, numt signs can be found in chromatograms when each sample was sequenced from both strands (Bensasson et al., 2001b; Leite, 2012). The ambiguities were taken as the numt sign at polymorphic sites in chromatograms of complementary sequences. Third, as each of F1-F5 represents the different length of COI and as these fragments partly or totally overlap with each other, the short sequences of the same sample compared with the larger/largest sequence (Fig. 2). The ambiguities between short and the longer/longest fragments of the same sample (in total 10 combinations between F1, F2, F2, F3, F4, and F5) were considered to be numt signs. After these comparisons, numt signs were searched in the contig sequences. Presence of stop codon in a contig produced from forward and reverse sequences of the same fragment and that of indel sites or any mutations in aligned F1-F5 sequences of the same individuals were considered as the numt signs. We also reconstructed the phylogenetic trees and checked the codon bias to distinguish numts in the contigs (Fig. 2). The sequences of the same fragment (F1-F5) arranged in various data matrices (for species of the same genus, genera of the same subfamily/family and the same suborder) and significantly divergent sequences were considered as numts (phylogenetic trees not presented). The mitochondrial genome is expected to be AT rich, therefore if the sequences show significant departure in AT/GC rate then it is also considered to be numts. The numt decision was mostly based on two or more of the above listed criteria.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

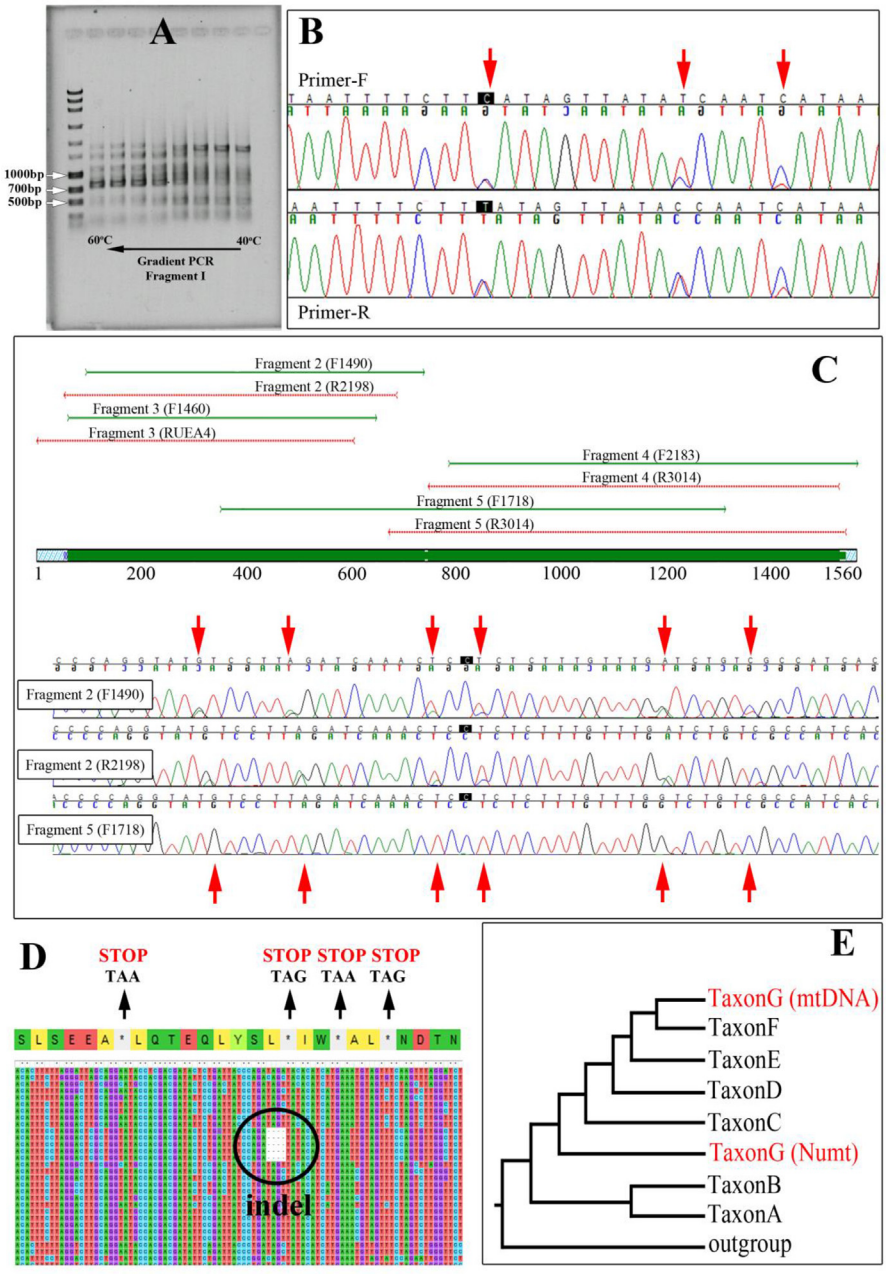

Fig. 2. Methods used to detect numts in PCR products, chromatograms or sequences. A) Multiple bands obtained by electrophoresis of PCR products, B) ambiguities at polymorphic sites/multiple piks in chromatograms of complementary sequences, C) The ambiguities between short and the longer/longest fragments of the same sample, D) Presence of stop codon, indel sites or non-synonymous mutations in contigs produced from forward and reverse sequences of the same fragment and that of in aligned F1-F5 sequences of the same sample, E) Unexpected phylogenetic location detected in trees produced from various data matrices (for species of the same genus, genera of the same subfamily/family and the same suborder).

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2.4. Data preparation and analyses

Rough complementary sequences were checked, edited and visually optimized using Sequencher v4.1.4 (Gene-Codes Corp.). The coding frames, the presence/absence of stop codons (TAA and TAG) and indels were checked by comparing sequences with the coding frame of *Drosophila* COI using DNASP v.5 (Librado and Rozas, 2009). The nucleotide sequences were translated into amino acid sequences using DNASP to detect non-synonymous substitutions. To establish matrices for the lineages and the sublineages (for species, genus, subfamily and family) sequences were aligned using MEGA v.5 (Tamura et al., 2013) with the MUSCLE algorithm (http://www. megasoftware.net), and phylogenetic analyses were conducted, by maximum likelihood algorithms in PAUP v.40b (Swofford, 2000) using appropriate substitution model that were calculated by jMODELTEST v.2 (Darriba et al., 2012). Descriptive parameters of the sequences such as nucleotide ratio, their frequencies per site and AT/GC ratio were calculated in ARLEQUIN v.3.01 (Excoffier vd., 2005) and DAMBE v.5.2.9 (Xia and Xie, 2001). Relative synonymous codon usage (Sharp et al., 1986), effective number of codons (Wright, 1990), codon bias index (Morton, 1993) and between group mean distance among sequence group were calculated in DNASP v. 5.0 and MEGA v.5.

Numt sequences resembling F4 found in genus *Anterastes* (Tettigoniidae) provided an opportunity to examine the phylogenetic position of numt sequences among non-numt sequences in a generic tree and compare divergence rate of numt and non-numt sequences. A data matrix was generated including six numt (this study) and 29 non-numt sequences of *Anterastes* representing all known species of the genus (29 of which previously uploaded to Genebank; Çıplak et al., 2010, 2015), and two outgroup sequences (one from each *Bolua turkiyae* and *Parapholidoptera distincta* (see Supporting Information). A BEAST analysis was conducted both to estimate origin plus divergence time of the numt sequences and to construct the phylogeny of haplotypes by Bayesian algorithms. BEAST was run using a Yule speciation process, uncorrelated lognormal relaxed clock and MCMC chains for 20×10^6 generations sampling every 1000^{th} generation. The convergence of stationary and the effective sample size of model parameters were checked using TRACER. The maximum clade-credibility trees were built with TREEANNOTATOR (Drummond and Rambaut, 2007), discarding the initial 10% of samples as burn-in. FIGTREE 1.3.1 (Rambaut, 2008) was used to visualize the results, including confidence intervals. The BEAST analysis was calibrated by the 0.0182 subs/s/Myr (see Kaya and Çıplak, 2016 for a discussion on this). The time to the most recent common ancestor (TMRCA) for each clade was estimated under the model parameters highlighted in jMODELTEST.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

3. Results

The PCR products amplified using forward/reverse primer couples of 1460/UEAI or UEA2/UEA2d targeting ~200 bp part of COI from 5′-3' upstream (F1) produced multiple bands in electrophoresis (Fig. 1) indicating co-amplification of non-mitochondrial COI or other experimental defects. Thus, no products to be sequenced were obtained to search for further signs of numts (Table 3).

F2 targeting ~700 bp of COI (the Folmer barcoding region) amplified using the forward/reverse primer couple of LCO1490/HCO2198 and PCR amplicons exhibited either multiple or a single band in electrophoresis. The amplicons with a single band (91 out of total 140 samples and 65% in Ensifera; 47 out of total 90 samples and 52.22% in Caelifera) were sent for sequencing to search further numt signs in chromatograms or their sequences. Of the 91single-banded amplicons of Ensifera 42 (46.15%) exhibited numt signs while 49 (53.85%) did not. In Caelifera, the number of chromatograms/sequences with numt signs were much higher as 39 of 47 single-band amplicons (82.98%) exhibited numt signs while only 8 (17.02%) did not (Table 3).

The number of PCR amplicons obtained using forward/reverse primer couple of UEA4/2191 targeting ~800 bp of the COI in 5′-3′ upstream (F3) producing a single-band in electrophoresis was extremely low especially in Ensifera (3.57%) but high in Caelifera (41.11%). The single-banded amplicons were sent for sequencing. Of the 5 chromatograms or sequences of Ensifera obtained from a single-band amplicons, 4 (80%) exhibited numt signs while only one did not. Of the 37 chromatograms or sequences of Caelifera obtained by sequencing the single-band amplicons 7 (18.92%) contained no numt signs while 30 (81.08%) exhibited one or more numt signs (Table 3).

The amplifications of F4 by forward/reverse primer couple of 2183/3034, which was targeting ~900 bp of COI in 5′-3′ upstream, produced 74 (52.86%) and 39 (43.33%) the single-banded electrophoretic amplicons in Ensifera and Caelifera respectively. Further numt signs were searched in chromatograms or in their sequences produced from the single-banded amplicons. The rate of chromatograms or sequences without numt signs was considerably increased by the fragment 4 in Ensifera (67 out of 74 (90.54%)). However, this number is still low in Caelifera (18 out of 39 (46.15%) while remaining 21 (53.85%) contains numt signs. The distinct result for F4 is the detection of a numt specific to *Anterastes* (Tettigoniidae) possibly obtained after the most recent common ancestor of the genus (see the last paragraph of this section). Of the six numt sequences three belong to *A. serbicus*, two to *A. burri* and one to *A. antitauricus* (Table 3).

The F5 targets ~1200 bp part of COI located in 5′-3′ upstream and amplified using primer couple of 1718/3014. The PCR products of F5 produced either multiple or

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

single (93 out of total 140 samples and 66.43% in Ensifera; 32 out of total 90 samples and 35.56% in Caelifera) bands in electrophoresis. The number of chromatograms or sequences without numt signs considerably high in Ensifera (within the 93 single-banded amplicons only 2 or 2.15% exhibited numt signs while 91or 97.85% did not). But, in Caelifera the number sequences with numt signs is still high (of the 32 single-band amplicons 11 (34.38%) exhibited numt signs while 21 (65.63%) did not (Table 3).

After alignment and trimming of the 6 numt and 29 non-numt sequences from *Anterastes*, the final length of the matrix was 743 bp. Of these 457 sites were constant, 286 were variable and 246 were parsimony informative, and each sequence constitutes a unique haplotype. Different phylogenetic analyses were applied to the data set (not shown here). In all analyses *Parapholidoptera distincta*, *Bolua turkiyae* and *A. disparalatus* branch off respectively in the base. The node after the *A. disparalatus* is unstable and mostly with a basal polytomy; (i) *A. antecessor*, (ii) the numt haplogroup, iii) *A. uludaghensis* + *A. davrazensis* and (iv) the haplogroup including a non-numt 22 haplotypes. The BEAST analyses were conducted in the light of these phylogenetic results to estimate gaining time of this numt. For a better time estimation and topology, BEAST constrained with two different phylogenetic options. The first constrain assumes that the paleonumts (see Song et al., 2014) were gained after branching off *A. disparalatus* leading to other members of the lineage (BEAST1). The second constrain assumes a later gaining of the numt as *A. disparalatus* + *A. antecessor* + the numt haplogroup + (*A. uludaghensis* + *A. davrazensis*) + the haplogroup including remaining 22 non-numt haplotypes (BEAST2). The non-numt haplotypes of *A. serbicus* and *A. antitauricus* placed in one of the crown groups consistent with intra-phylogeny of *Anterastes* (see Çıplak et al., 2015) while numt haplotypes constituted a basal clade within the genus after emergence of *A. disparalatus* (Fig. 3). Non-numt sequences obtained from the same samples placed within the crown group of *A. serbicus* + *A. antitauricus* clade (indicated by an * on the Fig. 3). The BEAST chronograms calibrated by the substitution rate for mitochondrial COI suggested 5.49 (BEAST1) and 4.91 (BEAST2) million years for the transition of ancestral numt from mtDNA to nDNA in respective BEAST analyses (Fig. 3). Since the evolutionary rate of the numts after transferring to nucleus will be different than the respective ancestral mitochondrial gene the later time estimations given on the BEAST chronogram will be misleading.

To compare divergence rates between numts and mitochondrial sequences we estimated three different between groups arithmetic mean of all pairwise distances (dxy); (i) the numt and non-numt sequences of the same species, (ii) numt sequences of different species, and (iii) non-numt sequences of different species (Table 4). The between group mean distances for numt versus non-numt sequences of the same species were 0.132 and 0.122 for *A. serbicus* and *A. antitauricus* respectively. The distance for numt sequences of *A. serbicus* and *A. antitauricus* was 0.018 while that for

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

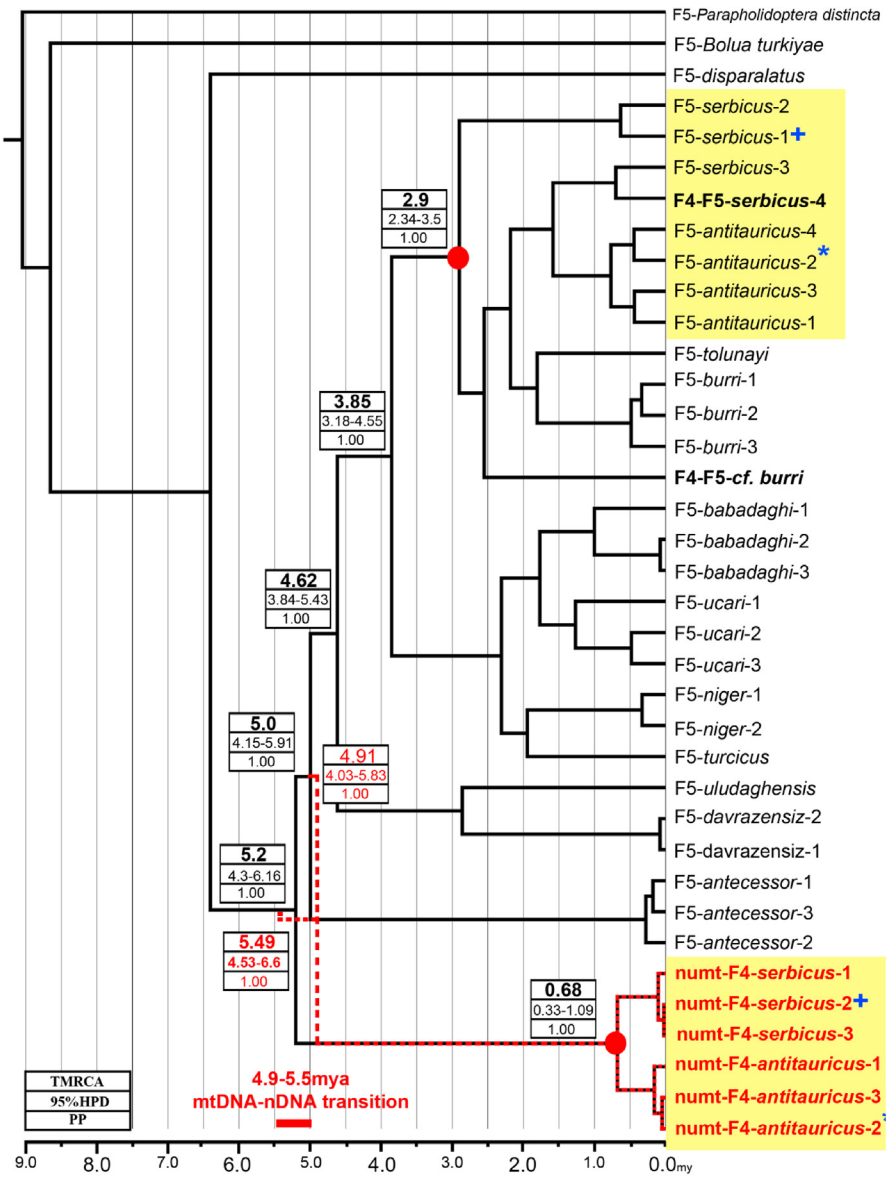

Fig. 3. The BEAST chronogram calculated using 6 numt (from F4; showed in red) and 29 non-numt (30 F5 sequences are given in Çıplak et al., 2015; + and * indicate to the numt and non-numt sequences from the same individual) sequences belonging to genus *Anterastes* (5.49 mya obtained from BEAST1 and 4.91 mya from BEAST2; see text for details) (TMRCA, time to most recent common ancestor; HPD, the highest posterior density; PP, posterior probability).

non-numt sequences of these two species was 0.061. The divergence time between non-numt *A. serbicus* and *A. antitauricus*is was estimated to be 2.9 million years ago but numt sequence of two clades was diverged 0.68 million years ago. The divergence rate of orthologous mtDNA sequence was 3.4−4.3 times higher than the numts sequences. These results indicate that mutation rate of numt genes (nuclear genome) is about $5.3-4.2 \times 10^{-9}$ per site/per generation/per million years.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 4. Between group mean distances (below diagonal) and SD for each (above diagonal) calculated using numt and non-numt mt-dna sequences of *A. serbicus* and *A. antitauricus*.

	numt-antitauricus	numt-serbicus	mt-antitauricus	mt- serbicus
numt-antitauricus		0,004	0.012	0.011
numt-serbicus	0.018		0.012	0.011
mt-antitauricus	0.122	0.124		0.006
mt-serbicus	0.130	0.132	0.061	

4. Discussion

Different data sets were produced from the samples studied. The number and the percentage of the samples with/without numt signs either can be calculated among the total number of specimens studied or among the total number of the PCR products sequenced. We could not obtain the single-banded PCR products to be sequenced by the primers targeting F1. Thus, there were no chromatograms or sequences to search for further numt signs. The main reason for these result is probably the length of this fragment COI (app. 200 bp) as numt co-amplification possibility decrease with the increase of targeted length (Bensasson et al., 2001b; Gaziev and Shaikhaev, 2010; Hazkani-Covo et al., 2010; Song et al., 2008, 2014; Richly and Leister, 2004). These results may indicate that numts of shorter regions of the respective marker are more common in all members of Orthoptera. It is well known that numts are not the only reason responsible from multiple electrophoretic bands, but, it is one of the major causes to be considered (Zang and Hewitt, 1996b; Kolokotronis et al., 2007; Triant and De Woody, 2007; Moulton et al., 2010). Data for other four fragments have further implications to be evaluated.

The number of sequences with and without numt signs shows a similar trend for all of F2, F3, F4 and F5 in all lineages either at the suborder or at the family level. The number of chromatograms/sequences without numt signs for F2, the fragment representing DNA barcode marker, is very high in all lineages, but, there is a considerable difference among them. The samples without numt signs constitute 46% in Ensifera, but it is only 17% in Caelifera (see Fig. 4A). According to rate of samples with numt signs, the families of Ensifera were ordered as Gryllotalpidae, Rhaphidophoridae, Tettigoniidae, and Gryllidae (data for Schizodactylidae are not reliable enough) while those of Caelifera as Pamphagidae, Acrididae and Pyrgomorphidae (see Fig. 4B). Unfortunately, the data do not allow an evaluation of F3 in Ensifera due to inappropriate primer selection and insufficient PCR amplifications, but a similar trend seems to be valid for the families belonging to Caelifera. The number of samples without numt signs considerably increases for the F4, representing ~900 bp of COI, in both suborders. The ninety percentages of chromatograms/sequences obtained from specimens of Ensifera could be reliably aligned and used for further

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

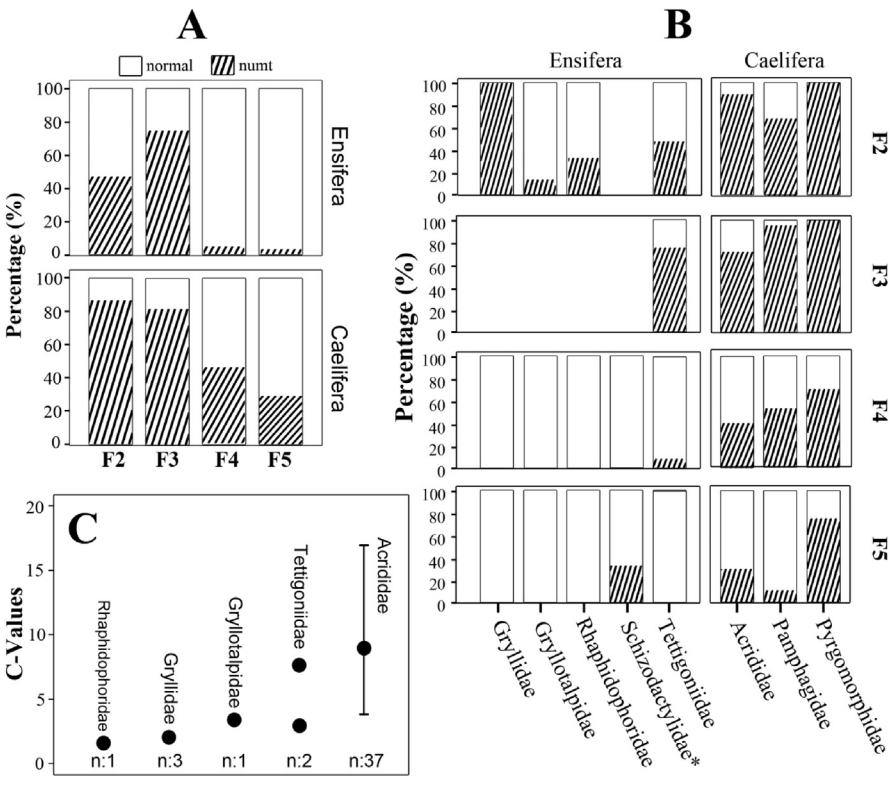

Fig. 4. The histograms showing the percentage of the chromatograms/sequences of F2-F5 with and without numt signs per suborder (A) and per family (B), and genome size for some families represented in this study (* inadequate sample size; n in C indicates the species number per family).

analyses. Only a few samples with numt signs were detected in Tettigoniidae while none in other four families of the suborder. But, results are contradictory in Caelifera as the percentage of samples with numt signs is still high. As for shorter fragments, the highest rate belonged to Pyrgomorphidae while there was a small difference between Acrididae and Pamphagidae. The longest fragment was F5 representing ~1200 bp part of COI. Among the total of 93 sequenced samples, there were only two with numt signs (in *Saga* and *Schizodatylus*) while remaining 91 (97%) were without numt signs. In contrary, the number of samples with numt signs was still very high (34%) in members of Caelifera, especially in Pyrgomorphidae. Of the remaining two families this rate was higher in Acrididae than in Pamphagidae.

In the light of these results it can be stated that numts are common in all members of Orthoptera as previously reported in several studies on the group (Bensasson et al., 2000, 2001a; Moulton et al., 2010; Song et al., 2008, 2014; Vaughan et al., 1999; Zang and Hewitt, 1996a). However, this statement is too much straightforward as the rate of sequences with numt signs considerably differs among lineages/sublineages in Orthoptera. Our data suggest that numt rate, independent of length, is higher in Caelifera than in Ensifera. There are also differences among the families of each suborder. Among the lineages within Ensifera, Tettigoniidae shows the highest rate while

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Pyrgomorphidae has the record among those of Caelifera. But, these rates may change depending on the length of fragments. For example, the rate of sequences with numt signs is higher in Acrididae than in Pamphagidae for F2 and F5 while the reverse is true for F3 and F4. A similar situation holds for members of Ensifera as the highest rate of sequences with numt signs detected in Gryllidae for F2 while no sequences with numt signs encountered for F4 and F5. In sum, numt prevalence is not similar even among members of Caelifera in contrary to previously suggested (Song et al., 2008, 2014) and the prevalence may change depending on the evolutionary origin of the paleonumt and length of the transferred segment (Song et al., 2014).

Our data indicate that there is a correlation between numt prevalence and genome size as suggested earlier for eukaryotes (Bensasson et al., 2001b; Gaziev and Shaikhaev, 2010; Hazkani-Covo et al., 2010; Richly and Leister, 2004). Although the haploid genome size record among animals is belonging to Orthoptera, there is a considerable variation among its members ranging from 1.51 to 16.56 Gb (1 C-value = 1pg = 978 Mbp) (Doležel et al., 2003). The C-value is known for 37 species of Acrididae (5.28−16.93; but mostly >10), 1 of Gryllotalpidae (3,339), 3 of Gryllidae (1,675-2.68), 1 of Rhaphidophoridae (1.55) and 2 of Tettigoniidae (2,592-3,03 or 7,61; the second value needs caution) while no data is available for Schizodactylidae, Pamphagidae and Pyrgomorphidae (Bensasson et al., 2001b; Hanrahan and Johnston, 2011; Song et al., 2014; Gregory, 2017) (Fig. 4C). Of the first five families, the highest records of genome size and the percentage of the sample with numt signs in our study are belonging to Acrididae. Although genome size is not known for Pamphagidae and Pyrgomorphidae numts are more prevalent in Caelifera and thus genome size and numt prevalence may show a similar tendency. When compared to Caelifera, the genome size is smaller in Ensifera and the number of sequences with numt signs for all fragments is accordingly lower. This makes the correlation between genome size and numt prevalence obvious (Fig. 4C). However, genome size data from members of Pamphagidae and Pyrgomorphidae, and from more members of Gryllidae, Rhaphidophoridae, and Tettigoniidae, will allow a better elucidation of this hypothesis.

Another main aim of present study was to test if there is a threshold between the length of mtDNA targeted to be amplified and a possibility of encountering numts. We studied five fragments (F1-F5) of COI and their lengths are ~200, ~700, ~800, ~900 and ~1200 bp respectively. The results to estimates a threshold for Ensifera and Caelifera are different. There are considerable numbers of sequences with numt signs for all of F1-F3 in both suborders. But, this number prominently reduces to <10% for F4 and to <3% for F5 in Ensifera while still considerably high in Caelifera. Thus a targeted length of ~1000 bp will provide more reliable sequences free of numts co-amplification in Ensifera. However, this threshold seems to be not very strict for Caelifera and some other methods to avoid numt co-amplification should be considered. Thus, this threshold should be kept in mind when the public sequences in databases downloaded and used.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

There are important implications for the barcoding fragment of COI. The F2 in this study amplified using the universal Folmer primer couple for insect (Folmer et al., 1994). By applying standard PCR protocol only 91out of total 140 samples in Ensifera and 47 out of total 90 samples being amplified and exhibit a single electrophoretic band as a reference for sequencing. Of the sequenced 91 amplicons of Ensifera 49 (54%) are without numt signs while 42 exhibit numt signs. This rate is much lower in Caelifera as the number of sequences without numt signs is 8 (17%) out of 47 sequenced samples. These results indicate unsuitability of Folmer region to be used for barcoding species and this case is more prominent for Caelifera. Similar results on shortcoming of this fragment for barcoding species have been reported especially for Orthoptera (Song et al., 2008; Moulton et al., 2010) and also for other animals (Sorenson and Quinn, 1998; Tautz et al., 2003; Rubinoff et al., 2006; Triant and De Woody, 2007; Buhay, 2009: Leite, 2012).

Analyses of the F4 paleonumts in genus *Anterastes* produced interesting results. Phylogenetic analyses showed that it gained in the ancestor of *Anterastes* after *A. disparalatus* diverged, and BEAST chronogram estimated its age as 5.49−4.91 million years. This means that these are pseudogenes and evolved as a part of nuclear genome after this date diverging at a different rate than the ancestral mitochondrial COI. The pairwise distances between numts (numt-antitauricus and numt-serbicus is 0.018), and numt and orthologous sequences of the same species (numt-antitauricus and mt-antitauricus as 0.122; numt-serbicus and mt-serbicus as 0.132) and between cross sequences (numt-antitauricus and mt-serbicus as 0.130; numt-serbicus and mt-antitauricus as 0.124) are similar and all these together are different than the pairwise distance between mt-antitauricus and mt-serbicus (as 0.061). These results support our numt assignment and indicate different divergence rate for numt and orthologous sequences. Our findings indicate that the divergence rate in mitochondrial COI is roughly four times rapid than its pseudogene migrated to the nucleus. It is a general expectation that mutation rate of pseudogenes is much faster than exon regions (Lopez et al., 1997; Bensasson et al., 2000, 2001a; Podlaha and Jianzhi, 2010). Accordingly, divergence rate in this study is slightly faster than nuclear genome mutation rate estimation of *Drosophila melanogaster* (3.5×10^{-9} per site per generation; Keightley et al., 2009) and also exon estimation (3.3×10^{-9} per site per generation) of Papadopoulou et al. (2010).

Declarations

Author contribution statement

Sarp Kaya: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Battal Çıplak: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by the Scientific and Technical Research Council of Turkey (Project no: TUBITAK-113T453) and by the Akdeniz University Research Fund, Turkey.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2018.e00929.

Acknowledgements

We thank Dr Hasan H. BAŞIBUYUK (Akdeniz University) for linguistically checking the manuscript.

References

Aljanabi, S.M., Martinez, I., 1997. Universal and rapid salt extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res. 25, 4692−4693. PMID: 9358185.

Antunes, A., Ramos, M.J., 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. Genomics 86, 708−717.

Avise, J.C., 2000. Phylogeography: the History and Formation of Species. Harvard University Press, Massachusetts.

Avise, J.C., 2009. Phylogeography: retrospect and prospect. J. Biogeogr. 36, 3−15.

Ballard, J.W.O., Rand, D.M., 2005. The population biology of mitochondrial DNA and its phylogenetic implications. Annu. Rev. Ecol. Evol. Syst. 36, 621−642.

Barrett, R.D.H., Hebert, P.D.N., 2005. Identifying spiders through DNA barcodes. Can. J. Zool. 83, 481−491.

Behura, S.K., 2007. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. Mol. Biol. Evol. 24 (7), 1492−1505.

2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Bensasson, D., Zhang, D.-X., Hewitt, G.M., 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. 17 (3), 406–415.

Bensasson, D., Petrov, D.A., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001a. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. 18 (2), 246–253.

Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001b. Mitochondrial pseudogenes: evolution's misplaced witnesses. TREE 16, 314–321. PMID: 11369110.

Bertheau, C., Schuler, H., Krumböck, S., Arthofer, W., Stauffer, C., 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol. Ecol. Resour. 11 (6), 1056–1059.

Berthier, K., Chapius, M.-P., Moosavi, S., Tohidi-Esefanahi, D., Sword, G.A., 2011. Nuclear insertions and heteroplasmy of mitochondrial DNA as two sources of intra-individual genomic variation in grasshoppers. Syst. Entomol. 36, 285–299.

Buhay, J.E., 2009. COI -like sequences are becoming problematic in molecular systematic and DNA barcoding studies. J. Crustac Biol. 29, 96–110.

Cao, G., Xue, R., Zhu, Y., Wei, Y., Gong, C., 2011. Nuclear mitochondrial DNA pseudogenes in the genome of the silkworm, *Bombyx mori*. J. Comput. Biol. Bioinf. Res. 3 (7), 103–119.

Chen, J., Li, Q., Kong, L., Yu, H., 2011. How DNA barcodes complement taxonomy and explore species diversity: the case study of a poorly understood marine fauna. PloS One 6 (6), e21326.

Çıplak, B., Kaya, S., Gundüz, I., 2010. Phylogeography of Anterastes serbicus species group (Orthoptera, Tettigoniidae): phylogroups correlate with mountain belts, but not with the morphospecies. J. Orthop. Res. JOR 19, 89–100.

Çıplak, B., Kaya, S., Boztepe, Z., Gündüz, I., 2015. Mountainous genus *Anterastes* (Orthoptera, Tettigoniidae): autochthonous survival in refugial habitats across several glacial ages via vertical range shifts. Zool. Scripta 44, 534–549.

Cristiano, M., Fernandes-Salomao, T., Yotoko, K., 2012. Nuclear mitochondrial DNA: an Achilles' heel of molecular systematics, phylogenetics and phylogeographic studies of stingless bees. Apidologie 43 (5), 527–538.

Dayama, G., Emery, S.B., Kidd, J.M., Mills, R.E., 2014. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 42 (20), 12640–12649.

 https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Meth. 9, 772.

Du, W.X., Qin, Y.C., 2015. Distribution of mitochondrial DNA fragments in the nuclear genome of the honeybee. Genet. Mol. Res. 14 (4), 13375−13379.

Doležel, J., Bartoš, J., Voglmayr, H., Greilhuber, J., 2003. Letter to the editor: nuclear DNA content and genome size of trout and human. Cytometry 51A (2), 127−128.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analyses by sampling trees. BMC Evol. Biol. 7, 214−221.

Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1, 47−50.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3 (5), 294−299.

Gaziev, A.I., Shaikhaev, G.O., 2010. Nuclear mitochondrial pseudogenes. Mol. Biol. (Moskow) 44, 405−417. PMID: 20608164.

Gellissen, G., Bradfield, J.Y., White, B.N., Wyatt, G.R., 1983. Mitochondrial DNA sequences in the nuclear genome of a locust. Nature 301, 631−634.

Gregory, T.R., 2017. Animal Genome Size Database. http://www.genomesize.com.

Hanrahan, S.J., Johnston, J.S., 2011. New genome size estimates of 134 species of arthropods. Chromosome Res. 19, 809−823.

Hazkani-Covo, E., Zeller, R.M., Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 6 (2), e1000834.

Hebert, P.D.N., Cywinska, A., Ball, S.L., de Waard, J.R., 2003a. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B Biol. Sci. 270, 313−321.

Hebert, P.D.N., Ratnasingham, S., de Waard, J.R., 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. Lond. B Biol. Sci. 270, S96−S99.

Hebert, P.D.N., de Waard, J.R., Landry, J.-F., 2010. DNA barcodes for 1/1000 of the animal kingdom. Bio. Lett. 6, 359−362.

Hubert, N., Hanner, R., Holm, E., Mandrak, N.E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J.,

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Bernatchez, L., 2008. Identifying Canadian Freshwater Fishes through DNA Barcodes. PloS One 3 (6), e2490.

Kaya, S., Çıplak, B., 2016. Budding speciation via peripheral isolation: the *Psorodonotus venosus* (Orthoptera, Tettigoniidae) species group example. Zool. Scripta 45, 521−537.

Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., Blaxter, M., 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome Res. 19, 1195−1201.

Kim, J.-H., Antunesa, A., Luoa, S.-J., Menninger, J., Nash, W.G., O'Briena, S.J., Johnson, W.E., 2006. Evolutionary analysis of a large mtDNA translocation (numt) into the nuclear genome of the *Panthera* genus species. Gene 366 (2), 292−302.

Kolokotronis, S.O., MacPhee, R.D.E., Greenwood, A.D., 2007. Detection of mitochondrial insertions in the nucleus (NuMts) of Pleistocene and modern muskoxen. BMC Evol. Biol. 7, 67.

Leite, L.A.R., 2012. Mitochondrial pseudogenes in insect DNA barcoding: differing points of view on the same issue. Biota Neotropica 12 (3), 301−308.

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451−1452.

Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39, 174−190. PMID: 7932781.

Lopez, J.V., Cevario, S., O'Brien, S.J., 1996. Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and a transposed mtDNA tandem repeat (*Numt*) in the nuclear genome. Genomics 33, 229−246.

Lopez, J.V., Culver, M., Stephens, J.C., Johnson, W.E., O'Brien, S.J., 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mol. Biol. Evol. 14 (3), 277−286.

Malik, A.N., Czajka, A., Cunningham, P., 2016. Accurate quantification of mouse mitochondrial DNA without co-amplification of nuclear mitochondrial insertion sequences. Mitochondrion 29, 59−64.

Moritz, C., Cicero, C., 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2 (10), e354.

Morton, B.R., 1993. Codon use and the rate of divergence of land plant chloroplast genes. Mol. Biol. Evol. 11, 231−238.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Moulton, M.J., Song, H., Michael, F.W., 2010. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). Mol. Ecol. Resour. 10, 615−627.

Papadopoulou, A., Anastasiou, I., Vogler, A.P., 2010. Revisiting the insect molecular clock: the Mid-Aegean Trench calibration. Mol. Biol. Evol. 27, 1659−1672.

Podlaha, O., Jianzhi, Z., 2010. Pseudogenes and their evolution. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd, Chichester.

Pons, J., Vogler, A.P., 2005. Complex pattern of coalescence and fast evolution of a mitochondrial rRNA pseudogene in a recent radiation of tiger beetles. Mol. Biol. Evol. 22 (4), 991−1000.

Rambaut, A., 2008. FIGTREE v1.2. Available via. http://tree.bio.ed.ac.uk/software/FigTree/.

Rawlings, T.A., MacInnis, M.J., Bieler, R., Boore, J.L., Collins, T.M., 2010. Sessile snails, dynamic genomes: gene rearrangements within the mitochondrial genome of a family of caenogastropod molluscs. BMC Genom. 11, 440.

Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21 (6), 1081−1084.

Rubinoff, D., Cameron, S., Will, K., 2006. A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J. Hered. 97 (6), 581−594.

Schmitz, J., Piskurek, O., Zischler, H., 2005. Forty million years of independent evolution: a mitochondrial gene and its corresponding nuclear pseudogene in primates. J. Mol. Evol. 61, 1−11.

Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R., 1986. Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125−5143. PMC311530.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., Flook, P., 1994. Evolution, weighting and the phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. 87, 651−701.

Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. PNAS, USA 105 (36), 13486−13491.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Song, H., Moulton, M.J., Hiatt, K.D., Whiting, M.F., 2013. Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: the biogeography of *Schistocerca* revisited. Cladistics 29, 643−662.

Song, H., Moulton, M.J., Whiting, M.F., 2014. Rampant nuclear insertion of mtDNA across diverse lineages within Orthoptera (Insecta). PloS One 9 (10), e110508.

Sorenson, M.D., Quinn, T.W., 1998. Numts: a challenge for avian systematics and population biology. Auk 115, 214−221.

Soto-Calderón, I.D., Lee, E.J., Jensen-Seaman, M.I., Anthony, N.M., 2012. Factors affecting the relative abundance of nuclear copies of mitochondrial DNA (Numts) in hominoids. J. Mol. Evol. 75 (3-4), 102−111.

Sun, X., Yang, A., 2016. Exceptionally large mitochondrial fragments to the nucleus in sequenced mollusk genomes. Mitochondrial DNA A 27 (2), 1409−1410.

Sunnucks, P., Hales, D.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. 13 (3), 51&524.

Swofford, D.L., 2000. PAUP*− Phylogenetic Analysis Using Parsimony (*and Other Methods). Ver. 4. [Computer Software and Manual]. Sinauer Associates, Sunderland, Massachusetts.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725−2729.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A.P., 2003. A plea for DNA taxonomy. TREE 18 (2), 70−74.

Triant, D.A., De Woody, J.A., 2007. The occurrence, detection, and avoidance of mitochondrial DNA translocations in mammalian systematics and phylogeography. J. Mammal. 88 (4), 908−920.

Triant, D.A., De Woody, J.A., 2008. Molecular analyses of mitochondrial pseudogenes within the nuclear genome of arvicoline rodents. Genetica 132 (1), 21−33.

Vaughan, H.E., Heslop-Harrison, J.S., Hewitt, G.M., 1999. The localization of mitochondrial sequences to chromosomal DNA in orthopterans. Genome 42, 874−880.

White, D.J., Wolf, J.N., Pierson, M., Gemmell, N.J., 2008. Revealing the hidden complexities of mtDNA inheritance. Mol. Ecol. 17, 4925−4942.

Williams, S.T., Knowlton, N., 2001. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. Mol. Biol. Evol. 18 (8), 1484−1493.

https://doi.org/10.1016/j.heliyon.2018.e00929
2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Wolff, J.N., Shearman, D.C.A., Brooks, R.C., Ballard, J.W.O., 2012. Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (numts). PloS One 7 (5), e37142.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23−29. PMID: 2110097.

Xia, X., Xie, Z., 2001. DAMBE: data analysis in molecular biology and evolution. J. Hered. 92, 371−373.

Zhang, D.X., Hewitt, G.M., 1996a. Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. Mol. Ecol. 5, 295−300.

Zhang, D.X., Hewitt, G.M., 1996b. Nuclear integrations: challenge for mitochondrial DNA markers. TREE 11, 247−251. PMID:21237827.

Zhang, D.X., Hewitt, G.M., 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12, 563−584.

 https://doi.org/10.1016/j.heliyon.2018.e00929

2405-8440/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).