Mitochondrial DNA, June–August 2010; 21(3–4): 87–104

informa
healthcare

Mitochondrial genomics in Orthoptera using MOSAS

NATHAN C. SHEFFIELD[1,2,*], KEVIN D. HIATT[1,*], MARK C. VALENTINE[1,3], HOJUN SONG[1], & MICHAEL F. WHITING[1]

[1]*Department of Biology, Brigham Young University, Provo, UT, USA,* [2]*Program in Computational Biology and Bioinformatics, Institute for Genome Sciences and Policy, Duke University, Durham, NC, USA, and* [3]*Department of Medicine, Howard Hughes Medical Institute, Washington University School of Medicine, St Louis, MO, USA*

(*Received 24 March 2010; revised 7 June 2010; accepted 9 June 2010*)

Abstract
We present complete mitochondrial genomes (mitogenomes) for three orthopterans (*Xyleus modestus*, *Physemacris variolosa*, and *Ellipes minuta*) and describe MOSAS (manipulation, organization, storage, and analysis of sequences), software we developed to facilitate annotation and analysis. We analyze the base composition, start and stop codons, non-coding regions, and gene order among these and 18 other orthopteran mitogenomes from GenBank and reconstruct a phylogeny of Orthoptera. We propose a tetranucleotide start codon for *cox1*, and hypothesize that the $tRNA^{Asp}$–$tRNA^{Lys}$ rearrangement is a synapomorphy for Acridomorpha, but not Caelifera. We further describe MOSAS, user-friendly software we used for this analysis. MOSAS streamlines sequence data storage, organization, annotation, and alignment, and provides convenient search tools for dataset construction and a robust annotation engine particularly suited to annotating mitogenomes (available at http://mosas.byu.edu).

Keywords: *Mitochondrial genome, Orthoptera, AT-bias, cox1 start codon, MOSAS*

Introduction

Metazoan mitochondrial genomes (mitogenomes) are composed of circular, double-stranded DNA and usually encode 37 genes in 15 kb or more of sequence. The genes typically encoded include tRNAs, rRNAs, and protein-coding genes from three protein complexes: Nicotinamide adenine dinucleotide (NADH) dehydrogenase, cytochrome oxidase, and ATP synthase. Mitogenome sequences (particularly coding regions) are commonly used as a source of phylogenetic data; furthermore, features such as gene order or base compositional bias are also useful phylogenetic markers (Boore and Brown 1998; Negrisolo et al. 2004; Podsiadlowski and Braband 2006).

The insect order Orthoptera (grasshoppers, katydids, and crickets) is the most diverse member of the group of polyneopteran insect orders, with more than 23,000 described species (Kevan 1982). Previous studies have primarily used combinations of nuclear and mitochondrial ribosomal genes to infer higher-level phylogeny of the order: 12*S* and 16*S* (Flook and Rowell 1997); 12*S*, 16*S*, and 18*S* (Flook et al. 1999); and 16*S*, 18*S*, and 28*S* (Jost and Shaw 2006). These strongly supported the monophyly of the two suborders, Caelifera and Ensifera; however, the family-level phylogeny of Orthoptera is yet to be fully resolved. Flook et al. (1995a) sequenced the first complete mitogenome from Orthoptera (*Locusta migratoria*) and reported a gene rearrangement ($tRNA^{Asp}$–$tRNA^{Lys}$) between the genes *cox2* and *atp8*, which was inverted from the inferred ancestral order, $tRNA^{Lys}$–$tRNA^{Asp}$ (Boore 1999). Flook et al. (1995b) then showed that this tRNA rearrangement was found in two members of the family Acrididae and one from Pyrgomorphidae, both of which belong to Caelifera. Fenn et al. (2008) also confirmed this pattern based on a larger taxon sampling from

Correspondence: N. C. Sheffield, Duke University, 101 Science Drive, Box 3382, Durham, NC 27708, USA. Tel: 1 919 684 2124. Fax: 1 919 668 4777. E-mail: nathan.sheffield@duke.edu
*N.C.S. and K.D.H. contributed equally.

ISSN 1940-1736 print/ISSN 1940-1744 online
DOI: 10.3109/19401736.2010.500812

Orthoptera. There are currently 20 complete orthopteran mitogenomes available on GenBank: 11 from Caelifera and 9 from Ensifera. However, these represent only two caeliferan and four ensiferan families. In this study, we present complete mitogenomes for three new caeliferan taxa from three unrepresented families. We use these three new mitogenomes along with those from GenBank to describe and compare the characteristics of the orthopteran mitogenomes. Herein, we describe orthopteran mitogenomes by analyzing the base composition, sequence divergence, start and stop codon usage, gene order, and non-coding DNA across all available complete mitogenomes for Orthoptera. We also reconstruct a preliminary phylogeny of Orthoptera based on mitogenome sequences.

As we have generated and analyzed a large number of complete mitogenomes, we have worked to streamline sequence storage, organization, annotation, and analysis. To this end, we have developed software to aid us in these tasks: MOSAS (manipulation, organization, storage, and analysis of sequences) which we have made openly available online (http://mosas.byu.edu). The development of MOSAS has been driven by the need for user-friendly bioinformatics tools. As it becomes easier and cheaper to generate sequence data, large datasets are becoming increasingly common. While this increase in availability is a boon to biologists, the scale of the resulting datasets requires more time to be spent storing, organizing, and analyzing data than ever before. Translating large amounts of sequence data into biologically meaningful and presentable information can require tedious hours of manual file manipulation and confusing processes at the command line. Given the volume of data available, there is an urgent need for user-friendly organizational tools that enable biologists to use these ever-growing datasets effectively. A few packages have been developed to address this challenge: for example, GARSA allows users to combine data from multiple sources (Davila et al. 2005), the Sequence Manipulation Suite performs some simple sequence format conversion and analysis on the web (Stothard 2000), and DOGMA annotates genomes automatically (Wyman et al. 2004). However, with the notable exception of Galaxy (Blankenberg et al. 2007), there are currently few software packages that attempt to streamline the entire process of storage, manipulation, organization, and analysis without requiring any programming knowledge, despite the fact that powerful tools such as BioPerl are freely available (Stajich et al. 2002). As such, we have developed MOSAS to fill this gap through a user-friendly web interface that allows users to employ powerful software without any programming experience. We used MOSAS in the analysis of orthopteran mitogenomes presented here.

Materials and methods

MOSAS software

MOSAS facilitates data organization by allowing users to store sequences in a personal database. Users can upload sequence files in FASTA format and link identifying information with each sequence. Sequences can also be added to the database through a number of tools that interface directly with GenBank to extract feature annotations or run BLAST. Once in the database, sequences and their identifying information are fully searchable. The search is a fast JavaScript/CSS implementation that allows a personal sequence database to be searched locally, preventing costly requests to the server. In our tests, a search of a database of over 1500 sequence names returned results in under 2 s. We have also implemented cutting-edge Ajax technology allowing users to make rapid in-place edits to text fields without reloading pages. Using these tools, users can separate sequences into meaningful subsets based on sequence name, organism, upload date, or user-defined notes. Once the sequences have been divided into datasets, users can analyze them in a variety of ways. We have installed a user-friendly interface to several alignment programs, including MUSCLE (Edgar 2004), MAFFT (Katoh et al. 2005), PROBCONS (Do et al. 2005), and INFERNAL (Dowell and Eddy 2006). Users can also calculate base composition statistics for an entire dataset. The datasets can be downloaded in a variety of formats via readseq, either before or after alignment (Gilbert 1999). Essentially, MOSAS provides a suite of tools that enables researchers to efficiently store, organize, and manipulate sequence data, all from a user-friendly web interface.

Taxon sampling

We analyzed entire mitogenomes from 21 taxa in this study, including the three caeliferan mitogenomes we sequenced: *Xyleus modestus* (Giglio-Tos 1894) (Romaleidae) (GenBank accession number GU945503), *Physemacris variolosa* (Linnaeus 1758) (Pneumoridae) (GenBank accession number GU945504), and *Ellipes minuta* (Scudder 1862) (Tridactylidae) (GenBank accession number GU945502). We included all of the mitogenomes for Orthoptera from GenBank except *L. migratoria migratoria* and *Gryllotalpa orientalis*, as these genera were represented twice (Table I). We chose to avoid including multiple representatives from the same genus because we primarily wanted to determine trends and relationships at the family level.

Mitogenome amplification and sequencing

We used primer walking to amplify and sequence three entire mitogenomes. We designed Orthoptera-specific primers and used them for the first round of PCR. These primers sit in *cox1*, *atp6*, *nad5*, *cytb*, *nad1*, *16S*,

Table I. Taxon sampling from Orthoptera used in the present study.

Suborder	Family	Genus/species	GenBank	Reference
Caelifera	Acrididae	*Schistocerca gregaria gregaria*	NC_013240	Erler et al. (2010)
		Oedaleus decorus asiaticus	NC_011115	Ma et al. (2009)
		Gastrimargus marmoratus	NC_011114	Ma et al. (2009)
		Chorthippus chinensis	NC_011095	Liu and Huang (2008)
		Oxya chinensis	NC_010219	Zhang and Huang (2008)
		Locusta migratoria	NC_001712	Flook et al. (1995a)
		Phlaeoba albonema	NC_011827	Shi et al. (2008)
		Calliptamus italicus	NC_011305	Fenn et al. (2008)
		Acrida willemsei	NC_011303	Fenn et al. (2008)
	Romaleidae	*Xyleus modestus*	GU945503	Present study
	Pyrgomorphidae	*Atractomorpha sinensis*	NC_011824	Ding et al. (2007)
	Pneumoridae	*Physemacris variolosa*	GU945504	Present study
	Tridactylidae	*Ellipes minuta*	GU945502	Present study
Ensifera	Tettigoniidae	*Anabrus simplex*	NC_009967	Fenn et al. (2007)
		Gampsocleis gratiosa	NC_011200	Zhou et al. (2008)
		Ruspolia dubia	NC_009876	Zhou et al. (2007)
		Deracantha onos	NC_011813	Zhou et al. (2009)
	Rhaphidophoridae	*Troglophilus neglectus*	NC_011306	Fenn et al. (2008)
	Gryllotalpidae	*Gryllotalpa pluvialis*	NC_011302	Fenn et al. (2008)
	Gryllidae	*Myrmecophilus manni*	NC_011301	Fenn et al. (2008)
		Teleogryllus emma	NC_011823	Ye et al. (2008)

and *12S*. A list of these Orthoptera-specific primers is available in Appendix 1. We used universal mitochondrial primers when specific primers did not work. Our long PCR amplicons spanned the following regions: *cox1–nad5* (~5.6 kb), *atp6–cytb* (~6.2 kb), *nad5–16S* (~4.4 kb), *cytb–12S* (~4 kb), *16S–cox1* (~4 kb), and *12S–cox2* (~4.5 kb). Long PCR helped to avoid coamplification of nuclear mitochondrial pseudogenes (numts) (Fenn et al. 2008). We ran nested PCR reactions on these products to obtain sequences of ~3 kb. After sequencing these nested products, we used them to design species-specific primers to amplify more of the genome until the entire mitogenome was complete. We sequenced double coverage or greater for both the major (J) and minor (N) strands to enable identification of potential numts. In cases where we suspected numt coamplification due to disagreement between sequences, we attempted to amplify regions outside the suspected numts with long PCR.

We used the Elongase PCR kit (Invitrogen, Carlsbad, CA, USA) for long and nested PCRs. We ran long PCRs under the following cycling protocol: 92°C for 2 min and 35 cycles of 92°C for 10 s, 50°C for 1 min, and 68°C for 7 min (increasing by 10 s each cycle for the last 25 cycles). Nested PCR conditions were 92°C for 2 min followed by 20 cycles of 92°C for 10 s, 50°C for 1 min, and 68°C for 4 min. When all the primers used in a reaction had melting temperatures greater than 56°C, we used an annealing temperature of 54°C instead of 50°C to increase the specificity of the PCR products. We performed sequencing reactions using ABI BigDye v3.1 dye terminating chemistry (Applied Biosystems, Inc., Foster City, CA, USA) and then fractionated them on an ABI 3730 capillary sequencer. We used the following cycling conditions for our sequencing reactions: 96°C for 1 min and 25 cycles of 96°C for 10 s, 50°C for 5 s, and 60°C for 75 s.

Annotating with MOSAS

One of the most powerful features of MOSAS is its annotation engine. After sequences have been uploaded, users can define features manually or use a collection of automated tools. Features generated using any method may be imported into the system through a universal import tool that reads a tab-delimited file of features. For automatic annotation, MOSAS provides a web interface to the program tRNAscan-SE (Lowe and Eddy 1997), which includes an updated covariance model for the elusive mitochondrial tRNA-SerAGN (Sheffield et al. 2008). Putative tRNA sequences found by tRNAscan-SE are added to the database as uncommitted features for later confirmation. MOSAS also provides an interface to ARWEN (Laslett and Canbäck 2008) to search for tRNA genes; users may choose between tRNAscan-SE and ARWEN. Users can visualize these tRNAs using RNAplot from the Vienna package (Hofacker et al. 1994) to inspect accuracy. A simple open reading frame (ORF) finder identifies possible gene boundaries. Users have the option of querying these ORFs in a BLAST search against a local database (of insect mitogenomes) to confirm genes by homology. After the ORF finder runs, users can manually inspect the boundaries to ensure correct annotation. To facilitate this process, we have implemented an annotation system designed to allow users to quickly and easily examine gene boundaries. This involves a "feature detail" screen that displays a feature with its flanking sequence. Amino acid

translations for all reading frames appear along the bottom of the page, and potential start or stop codons are highlighted in green or red. By clicking on a base in the sequence, users can change the starting or ending points of the feature. The program handles genes on the (−) strand seamlessly, appropriately taking the reverse complement of surrounding features. Also, rRNAs can be located initially using a sliding-window approach that queries a local database to find general locations of rRNAs, which can later be refined manually.

During annotation, users have access to a number of visualization tools. In addition to the feature detail view, a linear visualization screen shows an overview of features, their locations, and directions. There is also a circular genome visualization module that provides a user-interface to CGView (Stothard and Wishart 2005) to create a high-quality image. Users can export annotated sequences in Sequin format to facilitate GenBank submission. Users can also download an annotation table similar to those commonly published in mitogenome annotation papers (e.g. Sheffield et al. 2008). These annotations can include gene names, positions, lengths, anticodons, start and stop codons, and a column for intergenic spacers or overlaps. Additionally, we have made it easy to extract the features in FASTA format individually, in a group, or in subsets based on gene type.

Annotating orthopteran mitogenomes

To prepare our sequences for uploading into MOSAS, we first proofread and assembled raw sequence data into contigs using Sequencher™ version 4.9 (Gene Codes Corporation, Ann Arbor, MI, USA). We used chromatogram data to help resolve ambiguities and pinpoint potential numts. Long PCR allowed resolution of all ambiguous regions except for several cases of possible heteroplasmy. We uploaded entire mitogenome sequence files into MOSAS for annotation. We used tools available in MOSAS (see above) to annotate the mitogenomes (Table II) and to obtain a circular visualization of each mitogenome (i.e. Figure 1).

We used conventional insect mitochondrial start codons (methionine, isoleucine, and occasionally leucine) for all genes except *cox1* (Table III). We allowed gene overlap between adjacent genes but not between protein-coding genes and tRNAs. Our start and stop codon assignments attempt to minimize gaps between gene regions while also minimizing gene overlap. Our rationale for doing so is that mitochondrial evolution is generally considered to be under selective pressure toward size reduction (Schneider and Ebert 2004) and that gene overlap is not favorable (Nardi et al. 2003; Castro et al. 2006; Sheffield et al. 2008). When non-coding regions do not contribute to mitochondrial gene products, it is reasonable that evolution should favor the reduction of such regions. At times, reduction of non-coding intergenic spacers has led to gene overlap. However, overlap can lead to other complications, such as the issue of separating mRNA transcripts (Castro et al. 2006; Podsiadlowski and Braband 2006; Sheffield et al. 2008).

When a full stop codon (TAA or TAG) caused an overlap between a protein-coding gene and a tRNA, we annotated the gene with a partial stop codon (T or TA) instead (Table IV). Polyadenylation following transcription probably converts these partial stop codons into full stop codons (Ojala et al. 1981). We annotated with full stop codons, however, in overlapping protein-coding genes. One exception is the transition between *nad4* and *nad4l* in *Chorthippus chinensis*: no full stop codon exists in the ORF between tRNA-Thr and tRNA-His, so a partial stop codon was assigned based on alignment of *nad4* and *nad4l* across Orthoptera (Liu Y and Huang Y, unpublished work available in GenBank).

To determine a potential, conserved *cox1* start codon for Orthoptera, we used MEGA 4 (Tamura et al. 2007) to translate DNA sequences to amino acid sequences and then aligned those sequences using MUSCLE (Edgar 2004). We chose to assign *cox1* start codons as the first sequence conserved across Orthoptera following the end of tRNA-Tyr.

Analysis of mitogenome trends

We divided the mitogenome into several partitions and used MOSAS to compute AT-bias for each partition. The partitions were entire protein-coding sequences, the three codon positions of those coding sequences, the three gene categories of those coding sequences (NADH dehydrogenase subunits, cytochrome oxidase subunits, and ATP synthase subunits), combined ribosomal RNA sequences, combined transfer RNA sequences, and control regions (Table V). We compared AT-bias in each of these partitions across our taxa to investigate what may have contributed to the variations in bias we observed in the entire mitogenome sequences (Figure 2). We also used this information to compare the composition of each base across the taxa for each of the three codon positions (Figure 3).

To observe the relationship between AT-content and amino acid composition, we grouped the amino acids according to the AT-content of their codons (with AT-rich codons having an A or T in both the first and second codon positions, and GT-rich codons having a C or G in both positions). We then calculated the percentage of each amino acid made up of the entire amino acid sequence of each of the three gene categories using MEGA 4 (Tamura et al. 2007) and graphed those percentages (Figure 4).

Using MEGA 4 (Tamura et al. 2007), we translated protein-coding sequences to amino acids and then calculated the percentage divergence for both individual genes and gene regions. We used the *p*-distance

Table II. Nucleotide positions and anticodons for the three new caeliferan mitogenomes.

Gene	Strand	Anticodon	*Ellipes minuta*	*Xyleus modestus*	*Physemacris variolosa*
tRNA-I	+	GAU	1–64 (0)	1–66 (0)	1–65 (0)
tRNA-Q	–	UUG	66–134 (1)	70–138 (3)	63–132 (–3)
tRNA-M	+	CAU	136–203 (1)	139–207 (0)	134–201 (1)
nad2	+		204–1217 (0)	208–1228 (0)	202–1219 (0)
tRNA-W	+	UCA	1225–1293 (7)	1229–1296 (0)	1220–1284 (0)
tRNA-C	–	GCA	1286–1350 (–8)	1289–1352 (–8)	1277–1340 (–8)
tRNA-Y	–	GUA	1359–1423 (8)	1364–1431 (11)	1341–1404 (0)
cox1	+		1431–2957 (7)	1439–2963 (7)	1412–2936 (7)
tRNA-L	+	UAA	2963–3028 (5)	2964–3029 (0)	2937–3002 (0)
cox2	+		3031–3711 (2)	3030–3714 (0)	3007–3688 (4)
tRNA-D	+	GUC	3789–3855 (–1)	3715–3782 (0)	3689–3753 (0)
tRNA-K	+	CUU	3718–3789 (6)	3789–3858 (6)	3757–3826 (3)
atp8	+		3856–4017 (0)	3878–4039 (19)	3846–4004 (19)
atp6	+		4014–4688 (–4)	4036–4710 (–4)	4001–4675 (–4)
cox3	+		4693–5484 (4)	4712–5509 (1)	4679–5467 (3)
tRNA-G	+	UCC	5491–5555 (6)	5511–5576 (1)	5468–5533 (0)
nad3	+		5556–5907 (0)	5577–5930 (0)	5534–5885 (0)
tRNA-A	+	UGC	5908–5973 (0)	5932–5996 (1)	5886–5949 (0)
tRNA-R	+	UCG	5982–6045 (8)	6000–6065 (3)	5952–6015 (2)
tRNA-N	+	GUU	6043–6107 (–3)	6067–6131 (1)	6024–6087 (8)
tRNA-S	+	GCU	6106–6166 (–2)	6132–6198 (0)	6088–6154 (0)
tRNA-E	+	UUC	6179–6245 (12)	6204–6268 (5)	6156–6219 (1)
tRNA-F	–	GAA	6264–6328 (18)	6267–6330 (–2)	6221–6283 (1)
nad5	–		6332–8074 (3)	6331–8063 (0)	6284–8016 (0)
tRNA-H	–	GUG	8075–8141 (0)	8064–8128 (0)	8017–8081 (0)
nad4	–		8181–9533 (39)	8133–9455 (4)	8082–9402 (0)
nad4l	–		9527–9817 (–7)	9461–9754 (5)	9408–9698 (5)
tRNA-T	+	UGU	9825–9890 (7)	9757–9824 (2)	9701–9763 (2)
tRNA-P	–	UGG	9890–9954 (–1)	9827–9894 (2)	9764–9826 (0)
nad6	+		9957–10,475 (2)	9897–10,412 (2)	9828–10,346 (1)
cob	+		10,485–11,623 (9)	10,416–11,555 (3)	10,346–11,483 (–1)
tRNA-S	+	UGA	11,624–11,691 (0)	11,558–11,627 (2)	11,484–11,554 (0)
nad1	–		11,710–12,657 (18)	11,887–12,834 (259)	11,586–12,533 (31)
tRNA-L	–	UAG	12,658–12,723 (0)	12,835–12,900 (0)	12,534–12,597 (0)
rrnL	–		12,724–14,027 (0)	12,901–14,226 (0)	12,598–13,910 (0)
tRNA-V	–	UAC	14,028–14,092 (0)	14,227–14,297 (0)	13,911–13,978 (0)
rrnS	–		14,093–14,869 (0)	14,298–15,102 (0)	13,979–14,727 (0)
control	N/A		14,870–15,451 (0)	15,102–15,723 (0)	14,728–17,004 (0)

Note: The order of tRNAs between *cox2* and *atp8* for *E. minuta* is actually tRNA-K–tRNA-D. Numbers in parentheses represent the number of intergenic nucleotides before the beginning of the gene.

model of amino acid divergence to calculate all possible pairwise divergence percentages and then averaged those percentages across all 21 taxa to determine the mean divergence for a particular gene or gene region (Figure 5). We also calculated sequence divergence using ProtTest v. 2.4 (Abascal et al. 2005) to select an appropriate model of protein evolution for each protein-coding gene and then used the models selected to calculate divergence between sequences in MEGA 4 (Tamura et al. 2007).

To address potential start codons for *cox1*, we aligned a 57-bp segment spanning the transition between tRNA-Tyr and *cox1* for all 21 taxa (Figure 6). We performed the alignment based on amino acid sequences using MUSCLE (Edgar 2004). Finally, we aligned the non-coding nucleotides between tRNA-Ser and *nad1* (Figure 7) to analyze whether there was conservation in this region that could be associated with a function in mitochondrial transcription.

Sequence alignment and phylogenetic analysis

Using MOSAS, we organized protein-coding gene sequences from each of our 21 taxa and 2 outgroups from Dictyoptera (*Tamolanica tamolana* and *Periplaneta fuliginosa*) into datasets for each gene. We then translated these gene sequences into amino acids using MEGA 4 (Tamura et al. 2007). We used MUSCLE (Edgar 2004) (default settings) to align the amino acid sequences separately by gene. Next, we used MacClade (Maddison and Maddison 2003) to concatenate the aligned amino acid sequences into a single data matrix of a total of 3896 aligned amino acids.

We analyzed the concatenated matrix in a parsimony framework using TNT 1.1 (Goloboff et al. 2008) and performed a combination of sectorial search, ratchet, drifting, and tree fusing with 100 initial tree bisection and reconnection (TBR3) random additions, finding the minimum length 10

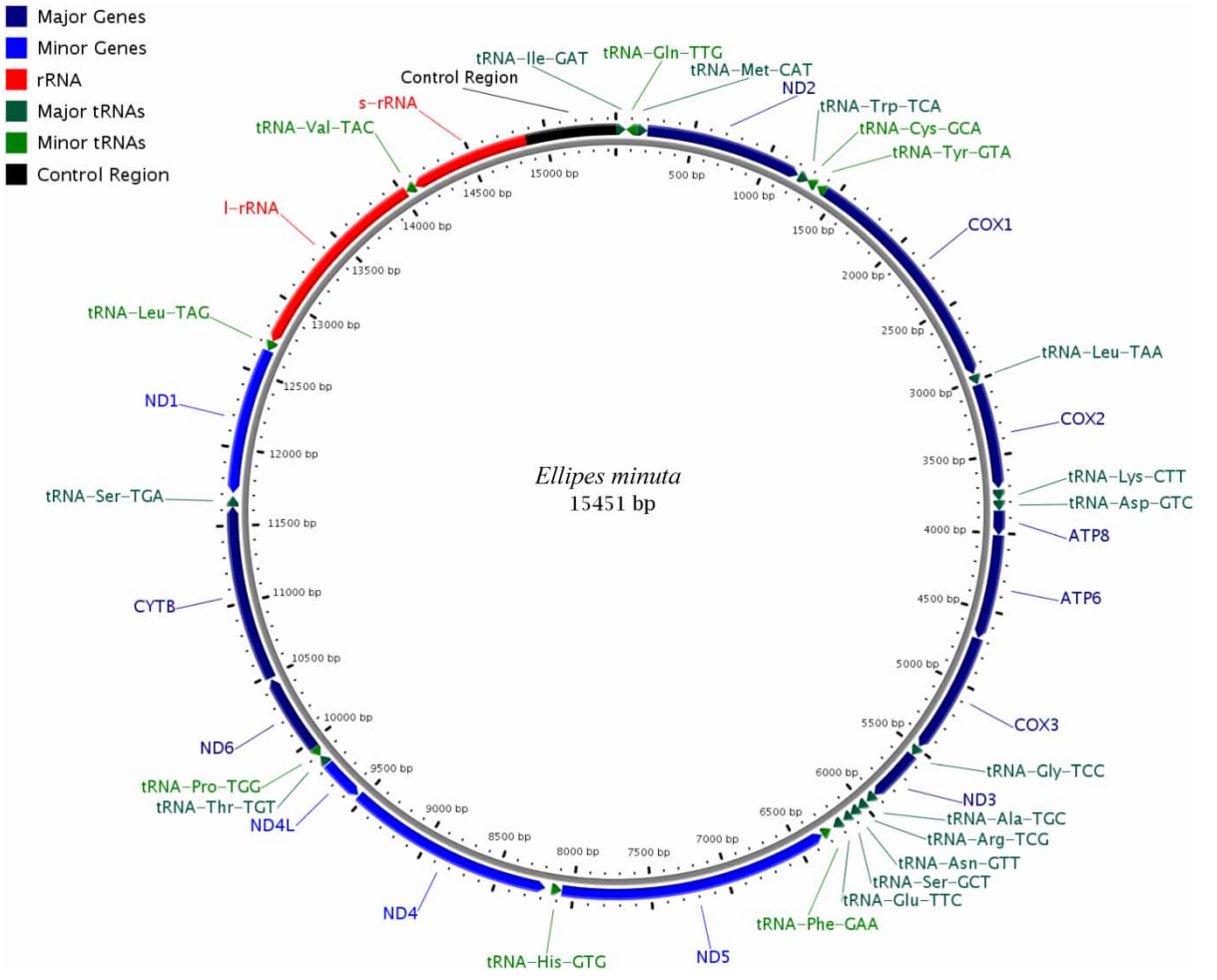

Figure 1. Circular visualization of the mitogenome of *E. minuta*.

times (Goloboff 1999; Nixon 1999). We calculated non-parametric bootstrap values using 1000 replicates with 100 TBR4 random additions.

Results and discussion

Mitogenome descriptions

We obtained complete mitogenome sequences from Romaleidae: *X. modestus* (15,723 bp), from Pneumoridae: *P. variolosa* (17,004 bp), and from Tridactylidae: *E. minuta* (15,451 bp). We selected these taxa to improve the representation of the diversity of Orthoptera with complete mitogenome sequences. *P. variolosa* had an unusually long control region (2277 bp), which makes it the longest orthopteran mitogenome to date compared with the mitogenomes available from GenBank. All three mitogenomes contained the same 37 genes (13 protein-coding genes, 2 rRNA genes, and 22 tRNA genes) typical to metazoan mitogenomes (Figure 1). They each followed the groundplan insect gene order with the exception of the order of $tRNA^{Lys}$ and $tRNA^{Asp}$ in *X. modestus* and *P. variolosa*, which was reversed from the traditional metazoan gene order (Boore 1999) (Table II).

Within these three mitogenomes, we discovered two unique, large non-coding regions other than the control region: one in the coding region of *X. modestus*, and the other within the control region of *P. variolosa*. *X. modestus* had a long spacer (259 bp) between tRNA-Ser and *nad1*. *P. variolosa* had a 343 bp sequence that repeated three times in tandem in the control region, contributing 1029 bp to the length of the region. Removing this repeat region would reduce the control region to 1248 bp, which is consistent with the lengths of control regions of other orthopterans. Analysis of the 259-bp intergenic spacer in *X. modestus* did not reveal any tandem repeats. Blastn searches of these two spacers did not return any strong clues concerning their potential origins.

We compared the three new caeliferan mitogenomes with 18 orthopteran mitogenomes from GenBank (Table I). Our analyses highlighted trends in AT-bias, uncovered non-coding spacers, compared start and stop codons, identified a potential *cox1* start codon for Orthoptera, and explored gene rearrangements. These findings and annotations reveal unique and common characteristics of mitogenomes in Orthoptera.

Table III. Start codon usage by gene across Orthoptera.

	nad2		*cox1*		*cox2*		*atp8*		*atp6*		*cox3*		*nad3*		*nad5*		*nad4*		*nad4l*		*nad6*		*cob*		*nad1*	
Genus	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon	AA	codon
Caelifera																										
Acrida	M	AUG	–	AUGA	M	AUG	M	AUA	M	AUA	M	AUG	I	AUU	M	AUA	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Calliptamus	M	AUG	–	AUGA	M	AUG	I	AUC	M	AUA	M	AUG	I	AUC	L	UUG	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Phlaeoba	I	AUC	–	AUGA	M	AUA	M	AUA	M	AUA	M	AUG	I	AUU	M	AUA	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Chorthippus	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUC	I	AUU	L	UUG	M	AUG	M	AUG	I	AUU	L	UUG
Locusta	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Gastrimargus	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Oedaleus	M	AUG	–	AUGA	M	AUG	I	AUC	M	AUA	M	AUG	I	AUU	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Oxya	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUC	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Schistocerca	M	AUG	–	AUGA	M	AUG	I	AUC	M	AUA	M	AUG	I	AUU	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	L	UUG
Xyleus	M	AUG	–	AUGA	I	AUU	I	AUC	M	AUA	M	AUA	I	AUC	L	UUG	L	UUG	M	AUG	I	AUC	M	AUG	L	UUG
Atractomorpha	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	I	AUU	M	AUG	M	AUG	M	AUG	M	AUG	I	AUU
Physemacris	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	M	AUG	M	AUA	M	AUG	I	AUC	M	AUG	L	UUG
Ellipes	I	AUU	–	AUGA	M	AUG	I	AUC	M	AUA	M	AUG	M	AUA	I	AUU	M	AUG	M	AUG	I	AUU	M	AUG	L	UUG
Ensifera																										
Gryllotalpa	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	M	AUG	M	AUA	M	AUG	I	AUC	M	AUG	M	AUG
Myrmecophilus	I	AUU	–	AUGA	M	AUA	I	AUC	M	AUA	M	AUG	I	AUU	L	UUG	I	AUU	M	AUG	I	AUU	M	AUG	I	AUU
Teleogryllus	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	I	AUU	I	AUU	M	AUG	I	AUU	M	AUG	L	UUG
Troglophilus	M	AUG	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUU	M	AUG	M	AUA	M	AUG	I	AUU	M	AUA	L	UUG
Anabrus	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	I	AUU	I	AUC	I	AUU	M	AUA	M	AUG	I	AUU	M	AUG	L	UUG
Gampsocleis	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	I	AUU	I	AUC	I	AUU	M	AUA	M	AUG	I	AUU	M	AUG	L	UUG
Ruspolia	I	AUU	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUA	I	AUU	I	AUU	M	AUG	M	AUG	M	AUA	M	AUA	I	AUC
	I	AUC	–	AUGA	M	AUG	I	AUU	M	AUA	M	AUG	I	AUC	I	AUU	M	AUA	M	AUG	I	AUU	I	AUU	L	UUG

Note: *cox1* start codons listed according to those assigned by the present study.

Table IV. Stop codon usage by gene across Orthoptera.

Genus	*nad2*	*cox1*	*cox2*	*atp8*	*atp6*	*cox3*	*nad3*	*nad5*	*nad4*	*nad4l*	*nad6*	*cob*	*nad1*
Caelifera													
Acrida	T	T	T	TAA	TAA	TAA	TAA	TA	TAG	TAA	TAA	TA	TAG
Calliptamus	T	T	TAA	TAA	TAA	TAA	TAA	TAG	TAA	TAA	TAA	TAA	TAG
Phlaeoba	TA	T	T	TAA	TAA	TAA	T	TA	TAA	TAA	TAA	TAA	TAG
Chorthippus	T	T	T	TAA	TAA	TAA	TAA	TAA	TAA	T	TAA	TAA	TAG
Locusta	T	TAG	T	TAA	TAA	TAA	T	T	TAA	TAA	TAA	TA	TAA
Gastrimargus	T	T	T	TAA	TAA	TAA	T	T	TAG	TAA	TAA	TA	TAA
Oedaleus	T	T	T	TAA	TAA	TAA	T	T	TAA	TAA	TAA	TA	TAA
Oxya	T	T	T	TAA	TAA	T	TA	T	TAA	TAA	TAA	T	TAG
Schistocerca	T	T	T	TAA	TAA	TAA	TAA	T	TAG	TAA	TAA	TAA	TAG
Xyleus	T	T	T	TAA	TAA	TAA	TAA	TA	TAG	TAA	TAA	TAA	TAG
Atractomorpha	TA	T	T	TAA	TAA	TAA	T	T	TAG	TAA	TAA	T	TAG
Physemacris	T	T	T	TAA	TAA	TAA	T	TA	T	TAA	TAA	T	TAG
Ellipes	TAA	TAA	TAA	TAG	TAA	TAA	T	TAA	TAG	TAA	TAA	TA	TAA
Ensifera													
Gryllotalpa	T	T	T	TAA	TAA	T	T	T	T	TAA	TAA	T	TAG
Myrmecophilus	T	T	T	TAA	TAA	TAA	T	T	T	TAA	TAA	TAA	TAG
Teleogryllus	T	T	T	TAA	TAA	TAG	TA	T	T	TAA	TAA	TA	TAA
Troglophilus	TA	T	T	TAA	TAA	TAA	T	TAG	TAA	TAA	TAA	T	TAG
Anabrus	T	T	T	TAA	TAA	TAA	T	T	T	TAA	TAA	T	TAA
Gampsocleis	T	T	T	TAA	TAG	TAA	TAA	T	T	TAA	TAA	T	TAG
Ruspolia	TAA	T	T	TAA	TAA	TA	TAA	T	TAA	TAA	TAA	TAG	TAA
Deracantha	T	T	T	TAA	TAA	TAA	T	T	T	TAA	TAA	T	TAA

Note: Partial stop codons designated with a T or TA.

They also allow an evaluation of MOSAS for mitogenome annotation and analysis.

AT-bias

Insect mitogenomes typically have higher percentages of adenine and thymine than they do of cytosine and guanine—a phenomenon known as AT-bias (Nardi et al. 2003; Castro et al. 2006; Cameron et al. 2007; Fenn et al. 2007; Cameron and Whiting 2008; Erler et al. 2010). This bias may be due in part to mitochondrial DNA damage that favors changes in nucleotide composition (Martin 1995). Mitochondrial DNA is susceptible to byproducts from respiration, including reactive oxygen species (such as superoxides: O_2^-). Reactive oxygen can denature sugar–base interactions, resulting in abasic sites. Randall et al. (1987) demonstrated that DNA polymerase favors inserting dATPs at abasic sites. Also, reactive oxygen species can alter bases, resulting in an increase in base mispairings. Cheng et al. (1992) identified a modification of guanine (to 8-hydroxyguanine) as a common mutation caused by reactive oxygen. This modification increases the chances of mispairings, often resulting in G to T substitutions (Cheng et al. 1992).

We observed varying degrees of AT-bias across Orthoptera (Table V). All taxa had AT-content greater than 50%; however, this bias was not distributed uniformly throughout the mitogenome (Figure 2). The coding region AT-content ranged from 65.3 to 75.5%. AT-bias was more extreme and had a wider range in the control region, with values ranging from 66.8 to 87.9%. Within Caelifera, AT-content ranged from 71.6 to 75.5% in the coding region for all taxa except *E. minuta*. This taxon had a coding region AT-content of only 66.3%. The control regions of caeliferan taxa had greater variance in AT-content, with values ranging from 81.2 to 87.9% for all taxa except *E. minuta*. Once again, this taxon had a much lower AT-content of 71.9%.

The AT-content was generally lower in Ensifera than in Caelifera. Within Ensifera, the AT-content ranged from 65.3 to 73.3% in the coding region and from 66.8 to 80.1% in the control region. A katydid, *Gampsocleis gratiosa* (Tettigoniidae), had the lowest overall AT-content.

Analysis of AT-bias based on codon positions demonstrated that third codon positions are more AT-rich than first and second codon positions. Second codon positions had the least variation (63.5–67.8% AT), while third codon positions had the greatest variation (66.3–91.3% AT) (Figure 2). These findings are consistent with other AT-content results in insects (Kim et al. 2006; Fenn et al. 2007, 2008; Zhou et al. 2007; Ma et al. 2009). The AT-content in *E. minuta* was lower than all other caeliferans at all three coding positions, but this difference was less extreme at the second position.

Furthermore, our analysis of base composition of protein-coding genes by codon position demonstrated that the composition of adenine and thymine is similar at first and third codon positions, whereas thymine is more abundant at second codon positions (Figure 3). This higher proportion of thymine is probably due to

Table V. Regional size and AT-bias in Orthoptera.

	Coding region		Codon position			Gene categories			Ribosomal RNAs		tRNAs		Control region
Taxon	Size	AT%	First	Second	Third	NAD	COX	ATP	Size	AT%	AT%	Size	AT%
Caelifera													
Acrida willemsei	14,818	75.5	69.9	66.1	90.2	77.6	70.2	77.7	2097	77.1	74.3	783	87.9
Calliptamus italicus	14,892	72.4	66.1	65.0	86.3	75.0	67.4	74.7	2123	73.7	70.5	783	86.7
Phlaeoba albonema	14,878	73.6	67.3	65.3	88.2	75.8	69.0	75.6	2109	74.9	71.7	779	83.3
Chorthippus chinensis	14,830	74.6	68.8	65.8	89.1	76.7	70.0	76.6	2108	76.0	72.4	769	83.9
Locusta migratoria	14,800	74.6	69.0	66.1	87.3	76.7	69.4	76.6	2094	77.6	74.0	922	85.9
Gastrimargus marmoratus	14,816	74.5	69.4	66.0	86.6	76.9	68.7	76.5	2106	77.0	74.8	1108	84.3
Oedaleus decorus asiaticus	14,813	74.4	68.7	65.7	87.4	76.6	69.1	76.9	2104	77.0	74.6	1446	84.4
Oxya chinensis	14,824	75.4	68.7	65.7	91.3	77.6	69.9	78.1	2109	77.5	73.3	619	87.4
Schistocerca gregaria gregaria	14,859	72.4	65.2	64.9	86.6	74.9	67.2	72.8	2131	74.0	71.2	766	87.1
Xyleus modestus	15,102	71.6	65.2	64.8	84.2	73.3	67.9	72.9	2131	73.3	70.8	621	84.9
Atractomorpha sinensis	14,737	73.9	69.5	66.3	85.1	75.1	69.9	77.2	2076	75.6	73.6	821	81.2
Physemacris variolosa	14,727	73.9	70.1	66.3	83.6	75.8	69.0	73.6	2062	76.0	75.4	2277	85.4
Ellipes minuta	14,869	66.3	61.1	63.5	70.5	67.1	62.3	63.1	2081	70.5	70.1	582	71.9
Ensifera													
Gryllotalpa pluvialis	14,603	71.8	65.1	64.8	83.6	73.5	67.5	71.0	1964	73.5	74.1	922	78.1
Myrmecophilus manni	14,534	70.0	64.1	64.7	78.3	71.0	65.2	70.5	1986	72.8	73.2	789	74.5
Teleogryllus emma	14,671	73.0	64.9	64.4	88.3	74.3	69.4	74.8	2056	73.9	74.8	989	73.9
Troglophilus neglectus	15,271	73.3	68.2	67.8	81.7	75.3	67.7	72.7	2127	75.5	75.4	539	66.8
Anabrus simplex	14,779	68.7	62.9	64.5	75.5	70.1	64.0	67.7	2097	71.4	73.1	987	80.1
Gampsocleis gratiosa	14,752	65.3	60.8	63.8	66.3	66.3	60.9	61.9	2086	69.4	71.5	1177	66.8
Ruspolia dubia	14,833	70.8	64.1	64.2	81.6	72.7	65.4	69.9	2116	73.2	73.4	138	77.5
Deracantha onos	14,760	68.7	63.5	64.3	75.6	70.3	64.7	67.0	2084	71.2	72.0	890	77.8

Note: Region lengths reported for the coding region and the two ribosomal RNA genes. AT-content presented for several regions of the mitogenome along with the three coding positions for protein-coding genes to analyze trends in AT-bias across the taxa.

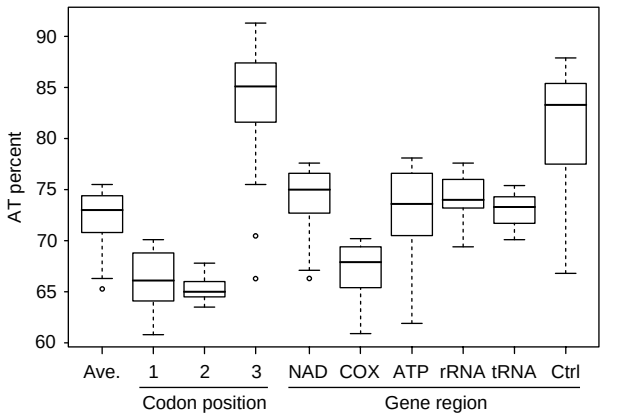

Figure 2. Boxplots summarizing AT-content among the taxa across different codon positions and different gene regions. Note: This figure corresponds to Table V.

the fact that leucine, isoleucine, and phenylalanine—three of the most abundant amino acids (see Figure 4)—all require a thymine at the second codon position. These findings are consistent with the base composition in beetles (Sheffield et al. 2008).

The average AT percentages across Orthoptera for the three gene categories were 73.9% for NADH dehydrogenase genes (NAD), 67.4% for cytochrome oxidase genes (COX), and 72.8% for ATP synthase genes (ATP). Without exception, AT-bias in the COX genes was lower than the other gene categories (Figure 2). This difference in AT-content between the gene categories may be attributable to a difference in amino acid composition between the categories (Figure 4). For five of the seven AT-rich amino acids, COX genes had a lower percentage than NAD or ATP genes, and for three of the four GC-rich amino acids, COX genes had a higher percentage than the other two gene categories. Therefore, the lower AT-content in the COX genes could be because they code for a higher percentage of amino acids that require GC-rich codons (Liu et al. 2008; Albu et al. 2009). This difference could also be due to a greater functional constraint on COX genes than on NAD or ATP genes (see the discussion on sequence divergence below).

Altogether, the least conserved portions of the mitogenome generally had the highest AT-content. For example, third codon positions, which are highly variable between taxa, exhibited higher AT-content than the other two coding positions. In the least constrained regions of the mitogenome, selection is weaker, so we would expect the nucleotide frequencies in these regions to strongly reflect the underlying patterns of mutation. Therefore, as mentioned above, the higher AT-content of these regions is possibly a reflection of the tendency of mtDNA evolution to favor AT-composition.

Sequence divergence

We observed that average divergence in the amino acid sequence of the 13 genes in the mitogenome ranged from 13% (in *cox1*) to 52% (in *nad6*) for our taxon sampling. We found COX genes to have lower sequence divergence than the NAD and ATP genes (Figure 5). Our model-based method (results not shown) confirmed the lower divergence of COX genes we observed in our *p*-distance analysis. This finding is consistent with an analysis done on substitution rates by gene (Cao et al. 1994; Pesole et al. 1999). All the three COX genes contribute to the functional region of complex IV of the electron transport chain. They are essential to the terminal delivery of electrons to oxygen, producing water, along with pumping protons across the inner mitochondrial membrane (Schultz and Chan 2001). Our finding suggests that changes in the amino acid sequences of COX genes are more likely to compromise the function of complex IV, than changes in NADH, or ATP genes are to impede the functions of complex I or complex V, respectively.

Our finding also confirms the connection we suggested previously between sequence divergence and AT-content. The most variable gene regions (NADH and ATP) had higher AT-content than

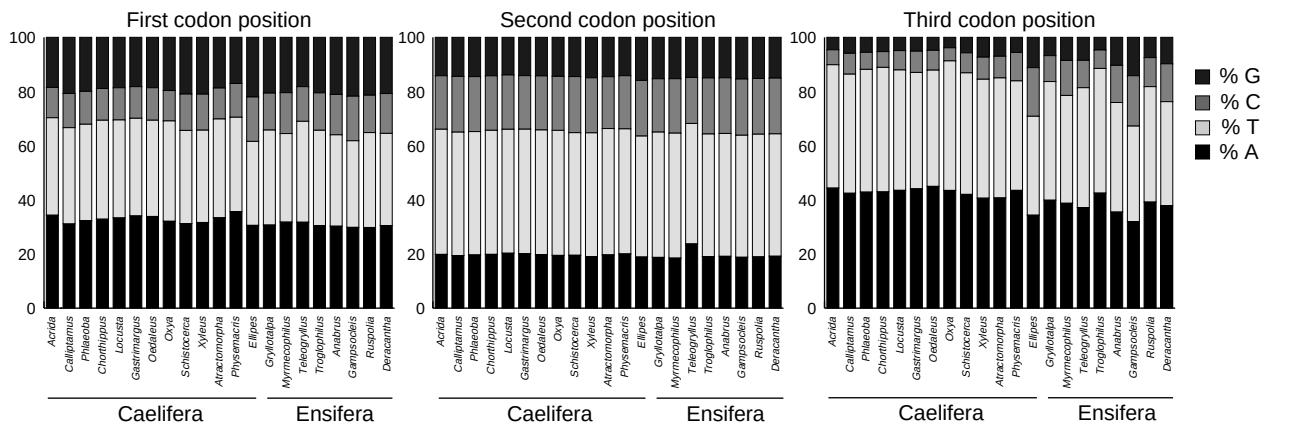

Figure 3. Base composition of all protein-coding genes. Note: Numbers of each nucleotide were converted to percentages and divided into a separate graph for each codon position.

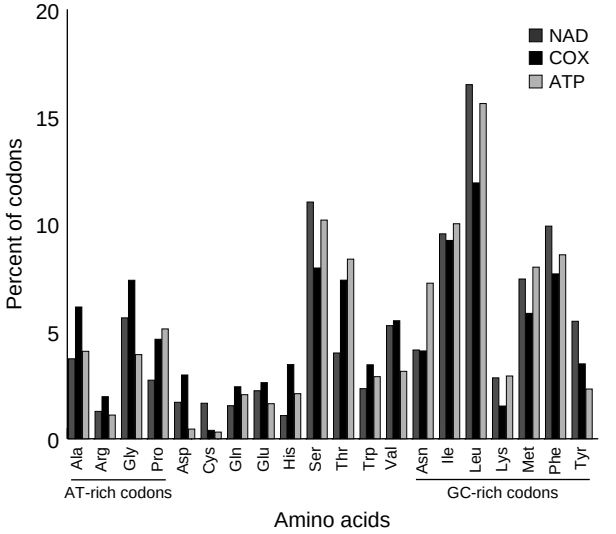

Figure 4. Amino acid usage by gene family. Note: Numbers of each amino acid were divided by gene region and then converted into percentages. AT-rich (WWN) and GC-rich (SSN) amino acids are identified.

the most conserved region (COX) within individual mitogenomes.

Non-coding DNA

Aside from the control region, most of the non-coding DNA we found interspersed in the mitogenomes we analyzed was not conserved in sequence or position and probably results from unique insertion events. However, we did find a non-coding spacer approximately 20 bp in length in all 21 taxa between tRNA-Ser and *nad1* (Figure 7). This intergenic spacer corresponds to spacers of similar length and genomic position found in Coleoptera (Sheffield et al. 2008) and Lepidoptera (Kim et al. 2006), and is conserved across our sample of taxa, with one exception: the annotation of *G. marmoratus* provided by Ma et al. (2009) has no spacer between tRNA-Ser and *nad1*. However, our alignment of the end of *nad1* across Caelifera suggests that, for this taxon, the gene should end 24 bp before the stop codon they annotated. Annotating a partial stop codon would maintain the presence of the conserved intergenic spacer found in all the other taxa (Figure 8).

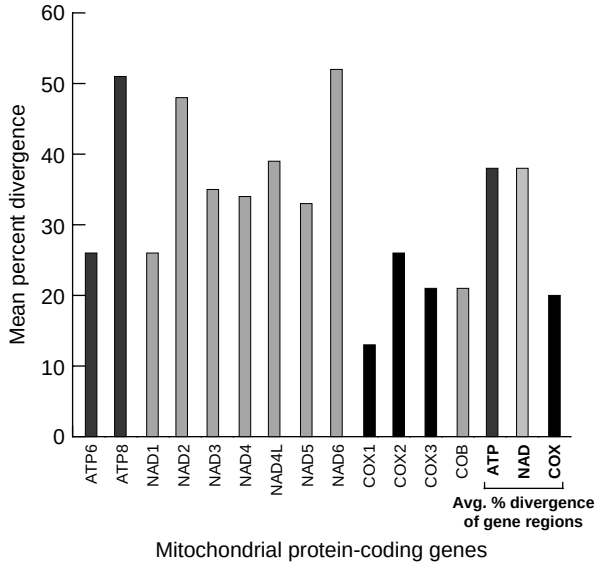

Figure 5. Mean percentage divergence in amino acid sequence of genes and gene regions across Orthoptera. Note: NAD genes are highlighted in light gray, ATP genes are highlighted in medium gray, and COX genes are highlighted in dark gray.

This spacer (Figure 7) may be conserved because this region functions as the binding site for a transcription-regulating peptide. One such peptide, DmTTF, has been demonstrated to bind between tRNA-Ser and *nad1* in *Drosophila melanogaster*, and binding of this peptide may signal termination of transcription of the major strand (Taanman 1999; Roberti et al. 2003). The conservation we observed between the aligned spacing regions of the 21 taxa supports the function of this region as the binding site for a similar transcription-terminating peptide.

Neither of the long non-coding regions in *X. modestus* nor *P. variolosa* appeared to have been formed by gene duplications, because neither contained long ORFs or folded as tRNAs. The finding of a tandem repeat in the control region of *P. variolosa* suggests that duplication of non-coding regions may contribute to the formation of large non-coding regions.

We observed several other intergenic spacers > 10 bp that were partially conserved. The most common of these were located between tRNA-Lys and atp8 (seven taxa from Caelifera, mean length of 16 bp), *nad5* and tRNA-His (six taxa from Caelifera and one from Ensifera, mean length of 14.7 bp), tRNA-Cys and tRNA-Tyr (five taxa from Caelifera, mean length of 10.8 bp), and tRNA-Gln and tRNA-Met (four taxa from Ensifera, mean length of 16 bp). Other spacers $>$ 10 bp were present, but not conserved in position across more than two taxa. We annotated a consistent overlap of 4 bp between *atp8* and *atp6* in each of the three caeliferan mitogenomes we generated.

Start and stop codons

Start codons were highly conserved for the majority of the protein-coding genes (Table III). The use of AUA (methionine) as a start codon for *atp6* and of AUG (methionine) for *nad4l* was completely conserved. Also, the start codons for *cox2*, *atp8*, *cox3*, *nad3*, *nad4*, *cob*, and *nad1* were mostly conserved. One exception was *nad5*, which employed a wide variety of start codons. Both *nad2* and *nad6* roughly favored AUU/C (isoleucine) as a start codon in Ensifera and AUG (methionine) as a start codon in Caelifera. Because these divisions are not definitive, the utility of start codons for these genes in phylogenetic inference is doubtful.

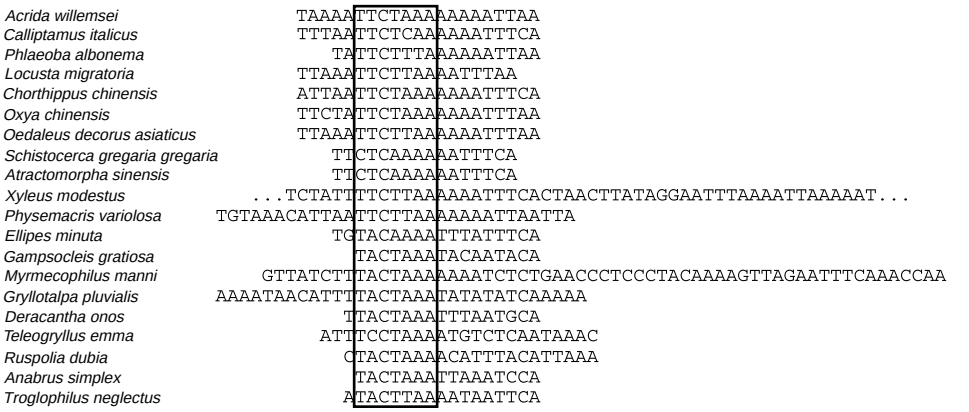

Figure 6. *cox1* start codon analysis based on an alignment of the region between tRNA-Tyr and *cox1*. Note: Previous annotations of *cox1* start codons along with our annotations in the three new mitogenomes are indicated with solid arrows. The nucleotides highlighted in gray represent the location of tRNA-Tyr. The four bases in bold and underlined indicate our proposed location of the *cox1* start codon for Orthoptera.

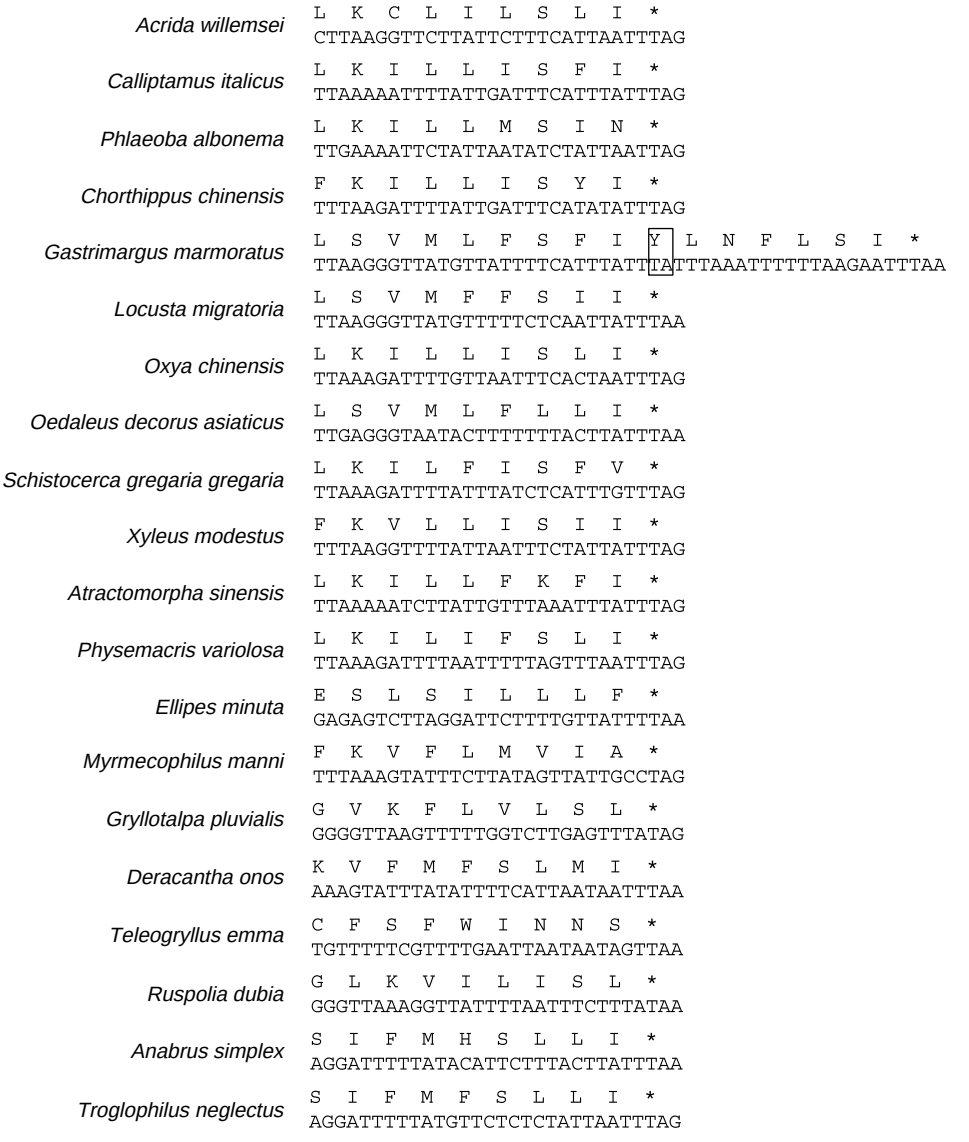

Figure 7. Alignment of non-coding spacer between tRNA-Ser and *nad1*. Note: The box indicates a conserved region within the spacer.

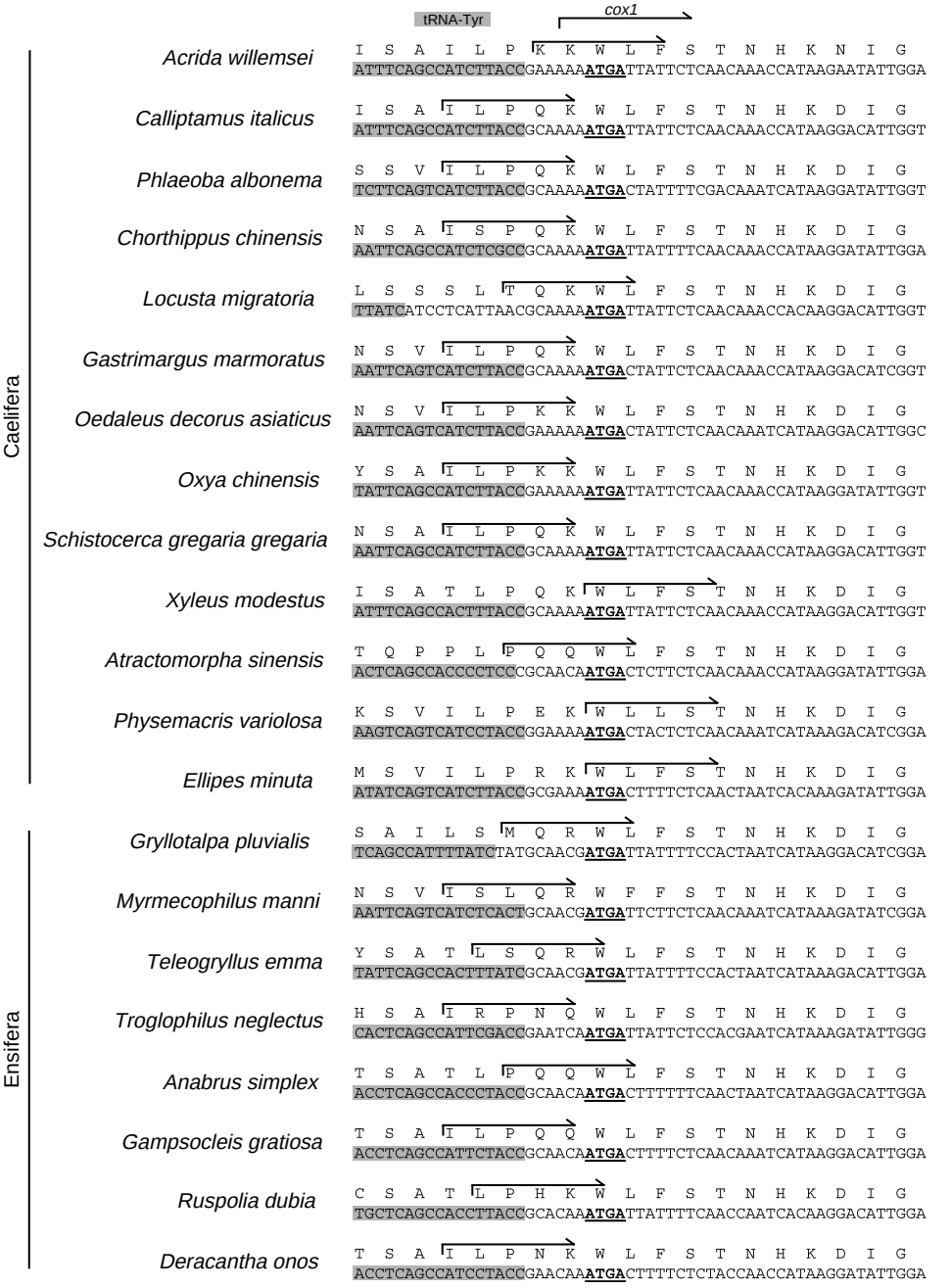

Figure 8. An alignment of the end of *nad1* across Orthoptera indicates that *Gastrimargus marmoratus* may use a partial stop codon. Note: Stop codons. The box around the TA in tyrosine indicates where we propose a partial stop codon is likely to be present based on the alignment.

We annotated partial stop codons for the majority of gene-tRNA interfaces (Table IV). However, three genes typically had full stop codons prior to the adjacent tRNA. These genes were *cox3*, *nad4*, and *nad1*. The non-coding intergenic spacer observed at the 3′ end of *nad1* led to the annotation of full stop codons, since a full stop codon always appeared well short of tRNA-Ser.

Cox1 start codons

Many have proposed the use of conserved, unconventional start codons for *cox1* because the ORF directly following tRNA-Tyr does not usually contain a conventional start codon within the first several amino acids in insects. Several start codons have been proposed, including nucleotide combinations coding for leucine, isoleucine, asparagine, glutamine, arginine, and serine (Nardi et al. 2003; Castro et al. 2006; Fenn et al. 2007; Sheffield et al. 2008; Stewart and Beckenbach 2009).

We decided upon a tetranucleotide start codon for *cox1* in Orthoptera: AUGA (Figure 6). This sequence was completely conserved across our taxon sampling and was the first completely conserved sequence

following the end of tRNA-Tyr. Similar tetranucleotide *cox1* start codons have been proposed for insects, including TTAA and ATAA in Drosophila (O'Clary and Wolstenholme 1983; Satta et al. 1987; Ballard 2000)—although a recent study has refuted this proposal, demonstrating that serine is probably the actual *cox1* start codon in *Drosophila* (Stewart and Beckenbach 2009)—and TTAG in Lepidoptera (Yukuhiro et al. 2002; Kim et al. 2006). O'Clary and Wolstenholme (1983) proposed ribosomal frameshifting as a cause of tetranucleotide start codons.

Using this tetranucleotide as a start codon for *cox1* generally introduces a 6-bp spacer between the 5′ end of tRNA-Tyr and the beginning of *cox1*. The majority of past annotations have selected isoleucine as the start codon for *cox1*, which normally requires an 8-bp overlap between the two regions. We doubt this annotation because of the protein-coding gene–tRNA overlap it introduces. Also, our observation that the sequence leading up to the codon AUGA is less conserved in the nucleotide and amino acid sequence than the sequence following the codon suggests that the gene's 5′ end occurs at this codon. However, our annotations will remain speculative until mRNA transcripts are analyzed to determine the actual start codon for *cox1*.

Gene order

The inversion from $tRNA^{Lys}$–$tRNA^{Asp}$ to $tRNA^{Asp}$–$tRNA^{Lys}$ has been considered a synapomorphy for Caelifera (Flook et al. 1995b; Fenn et al. 2008). However, one of the 13 caeliferan mitogenomes—the tridactylid *E. minuta*—maintained the ancestral gene order without the tRNA rearrangement. Tridactylidae, which is commonly known as the pygmy mole cricket, is the most basal lineage of Caelifera, belonging to its own superfamily—Tridactyloidea (Flook et al. 1999). Our finding suggests that the rearrangement occurred after the divergence of the tridactylid lineage and is, therefore, not a synapomorphy for Caelifera (Flook et al. 1995b; Fenn et al. 2008). However, we did find that the rearrangement to $tRNA^{Asp}$–$tRNA^{Lys}$ is conserved across the caeliferan group Acridomorpha, and, therefore, propose that it is a synapomorphy for that clade (Figure 6).

A different gene rearrangement has been reported uniquely in a cricket, *Teleogryllus emma*. Ye et al. (2008) discovered that tRNA-Asn, tRNA-Ser, and tRNA-Glu, which are normally coded on the major strand (Boore 1999), are coded on the minor strand in *T. emma*. They noted that this transposition resulted in an inversion of the gene order, changing order to $tRNA^{Glu}$–$tRNA^{Ser}$–$tRNA^{Asn}$ (Ye et al. 2008) (Figure 9). This rearrangement appears to be unique within Orthoptera and has not been observed in any other ensiferan species that are phylogenetically close to Gryllidae.

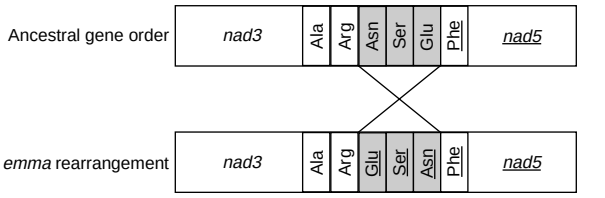

Figure 9. tRNA rearrangement unique to *T. emma*. Note: In this taxon, tRNA-Asn, tRNA-Ser, and tRNA-Glu are coded on the minor instead of the major strand, and their order is inverted from the ancestral gene order. Genes coded on the minor strand are underlined.

Phylogenetic analysis

Our parsimony analysis of the amino acid sequences of protein-coding genes recovered a single most parsimonious tree (length = 12,712, CI = 0.62, RI = 0.55) (Figure 10). This preliminary topology is largely congruent with previous phylogenetic hypotheses of Orthoptera based on molecular (Flook et al. 1999; Jost and Shaw 2006; Fenn et al. 2008; Erler et al. 2010) and morphological (Eades 2000) datasets. Our analysis recovers the monophyly of Orthoptera, Caelifera, and Ensifera. The monophyly of these groups is not well supported, but may gain greater support as we obtain a denser taxon sampling. Furthermore, it identifies the family Tridactylidae (*E. minuta*) as the most basal of the caeliferan taxa we included. *X. modestus*, our representative from the family Romaleidae, nests within the family Acrididae, rendering that family paraphyletic. A denser taxonomic sampling is also needed to address whether Acrididae is a monophyletic group.

Other than the tRNA rearrangement that is a molecular synapomorphy for Acridomorpha (Figure 10), there are no observable phylogenetic trends in base compositional bias, start and stop codon usage in protein-coding genes, or in the secondary structures and anticodons of tRNAs. The overall AT percentages of the members of Tettigoniidae and Acrididae appear to be generally low and high, respectively, but this pattern is not conclusive. The conservation of the genome structure is remarkable in an ancient lineage such as Orthoptera, compared with highly variable genome structures known in Hymenoptera (Dowton and Austin 1999).

Evaluation of MOSAS

We used MOSAS to annotate the other 18 mitogenomes included in the present study and compared those annotations with the annotations available on GenBank. We found the MOSAS software especially helpful in avoiding annotation errors and determining start and stop codons of protein-coding genes. For example, we consistently annotated a 4-bp overlap between *atp8* and *atp6* in contrast to the 7-bp overlap required by 15 of the 18 GenBank annotations because they chose a start

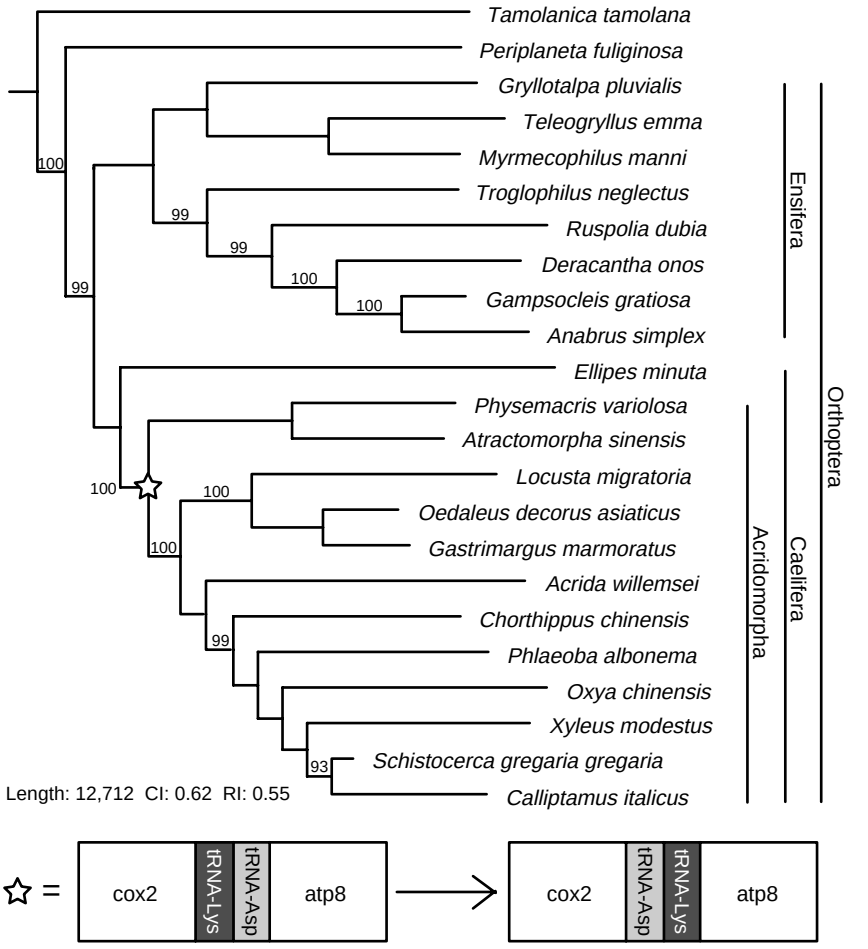

Figure 10. Maximum parsimony phylogeny of Orthoptera based on amino acid sequences of protein-coding genes. Note: *T. tamolana* and *P. fuliginosa* are outgroup taxa from Dictyoptera. We recovered one most parsimonious tree and reported bootstrap values above 90. The star indicates the derivation of the tRNA rearrangement from $tRNA^{Lys}$–$tRNA^{Asp}$ to $tRNA^{Asp}$–$tRNA^{Lys}$ in the evolution of Orthoptera. This rearrangement is a synapomorphy for Acridomorpha.

codon upstream from the one we chose. Also, 15 of the 18 annotations from GenBank did not consider UUG (Leu) as a possible start codon for *nad1*, which led to the annotation of unnecessarily long non-coding gaps between tRNA-Leu and *nad1*. In many other cases, not recognizing isoleucine as a plausible start codon led to large non-coding gaps between gene regions. Lastly, the GenBank annotations of *Acrida willemsei* and *Calliptamus italicus* used partial stop codons between *atp6* and *cox3*, although full stop codons were present and did not introduce overlap between the two genes. We annotated full stop codons in both cases to minimize intergenic spacing. Altogether, annotating with MOSAS allowed for quick, consistent annotations.

We have also used MOSAS for a variety of other projects, including the concatenation and partitioning of alignments for input into phylogenetic software. For mitogenome annotation, MOSAS is similar to DOGMA; however, we have found MOSAS superior in our research for several reasons, including access to tRNAscan-SE and ARWEN for tRNA finding, improved feature detail display (particularly for (–) strand genes), better import and output formats, better RNA visualization via RNAplot, and faster database updates with Ajax programming. We believe that MOSAS will be useful to anyone whose work involves the assembly of (moderately) large sets of sequence data, particularly for those interested in mitochondrial genomics. Workers with datasets containing fewer than 2000 sequences are encouraged to use our web resources to help with data organization. For larger datasets, workers may download the source code to set up a local server. A comprehensive manual that describes how to perform common tasks is available as a PDF file on the web site.

Acknowledgments

Funding for this research was provided by National Science Foundation grants EF-0531665 to M.F.W. and DEB-0816962 to H.S. and M.F.W., and a Brigham Young University Office of Research and Creative Activities grant to N.C.S.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Albu M, Min XJ, Golding GB, Hickey D. 2009. Nucleotide substitution bias within the genus Drosophila affects the pattern of proteome evolution. Genome Biol Evol 1:288–293.

Ballard JWO. 2000. Comparative genomics of mitochondrial DNA in members of the Drosophila melanogaster subgroup. J Mol Evol 51:48–63.

Blankenberg D, Taylor J, Schenck I, He JB, Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller W, Makova KD, Hardison RC, Nekrutenko1 A. 2007. A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. Genome Res 17:960–964.

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780.

Boore JL, Brown WM. 1998. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8:668–674.

Cameron SL, Whiting MF. 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.

Cameron SL, Lambkin CL, Barker SC, Whiting MF. 2007. A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. Syst Entomol 32:40–59.

Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M. 1994. Phylogenetic relationships among Eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J Mol Evol 39:519–527.

Castro LR, Ruberu K, Dowton M. 2006. Mitochondrial genomes of *Vanhornia eucnemidarum* (Apocrita: Vanhorniidae) and *Primeuchroeus* spp. (Aculeata: Chrysididae): Evidence of rearranged mitochondrial genomes within the Apocrita (Insecta: Hymenoptera). Genome 49:752–766.

Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G-]T and a-]C substitutions. J Biol Chem 267: 166–172.

Davila AMR, Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, Mazzoni CJ, Wagner G, Pires PF, Grisard EC, Cavalcanti MCR, Campos MLM. 2005. GARSA: Genomic analysis resources for sequence annotation. Bioinformatics 21: 4302–4303.

Ding F-M, Shi H-W, Huang Y. 2007. Complete mitochondrial genome and secondary structures of lrRNA and srRNA of Atractomorpha sinensis (Orthoptera, Pyrgomorphidae). Zoological Research 29:580–588.

Dowell RD, Eddy SR. 2006. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinform 7:400.

Dowton M, Austin AD. 1999. Evolutionary dynamics of a mitochondrial rearrangement "hot spot" in the hymenoptera. Mol Biol Evol 16:298–309.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340.

Eades DC. 2000. Evolutionary relationships of phallic structures of Acridomorpha (Orthoptera). J Orthoptera Res 9:181–210.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

Erler S, Ferenz HJ, Moritz RFA, Kaatz HH. 2010. Analysis of the mitochondrial genome of *Schistocerca gregaria gregaria* (Orthoptera: Acrididae). Biol J Linnean Soc 99:296–305.

Fenn JD, Cameron SL, Whiting MF. 2007. The complete mitochondrial genome sequence of the Mormon cricket (*Anabrus simplex*: Tettigoniidae: Orthoptera) and an analysis of control region variability. Insect Mol Biol 16:239–252.

Fenn JD, Song H, Cameron SL, Whiting MF. 2008. A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. Mol Phylogenet Evol 49:59–68.

Flook PK, Rowell CHF. 1997. The phylogeny of the Caelifera (Insecta, Orthoptera) as deduced from mtrRNA gene sequences. Mol Phylogenet Evol 8:89–103.

Flook PK, Rowell CHF, Gellissen G. 1995a. The sequence, organization, and evolution of the Locusta migratoria mitochondrial genome. J Mol Evol 41:928–941.

Flook PK, Rowell CHF, Gellissen G. 1995b. Homoplastic rearrangements of insect mitochondrial transfer-RNA genes. Naturwissenschaften 82:336–337.

Flook PK, Klee S, Rowell CHF. 1999. Combined molecular phylogenetic analysis of the Orthoptera (Arthropoda, Insecta) and implications for their higher systematics. Syst Biol 48: 233–253.

Gilbert DG. 1999. ReadSeq version 2, an improved biosequence conversion tool, written in the Java language. Bloomington, IN: Bionet. Software. Available from: http://iubio.bio.indiana.edu/soft/molbio/readseq/java [accessed March 2010].

Goloboff PA. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. Cladistics 15(4):415–428.

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24:1–13.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. Monatshefte Chem 125:167–188.

Jost MC, Shaw KL. 2006. Phylogeny of Ensifera (Hexapoda: Orthoptera) using three ribosomal loci, with implications for the evolution of acoustic communication. Mol Phylogenet Evol 38: 510–530.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518.

Kevan DKM. 1982. Orthoptera. In: Parker SP, editor. Synopsis and classification of living organisms. New York: McGraw-Hill. p 352–383.

Kim I, Lee EM, Seol KY, Yun EY, Lee YB, Hwang JS, Jin BR. 2006. The mitochondrial genome of the Korean hairstreak, *Coreana raphaelis* (Lepidoptera: Lycaenidae). Insect Mol Biol 15: 217–225.

Laslett D, Canbäck B. 2008. ARWEN: A program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24:172–175.

Liu Y, Huang Y. 2008. Sequencing and analysis of complete mitochondrial genome of Chorthippus chinensis Tarb. Chinese Journal of Biochemistry and Molecular Biology 24:329–335.

Liu Y, Li Y, Pan M, Dai F, Zhu X, Lu C, Xiang Z. 2008. The complete mitochondrial genome of the Chinese oak silkmoth, *Antheraea pernyi* (Lepidoptera: Saturniidae). Acta Biochim Biophys Sin 40:693–703.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964.

Maddison WP, Maddison DR. 2003. MacClade version 4.06. Sunderland, MA: Sinauer Associates. Available from: http://macclade.org.

Martin AP. 1995. Metabolic-rate and directional nucleotide substitution in animal mitochondrial-DNA. Mol Biol Evol 12: 1124–1131.

Ma C, Liu CX, Yang PC, Kang L. 2009. The complete mitochondrial genomes of two band-winged grasshoppers, *Gastrimargus marmoratus* and *Oedaleus asiaticus*. BMC Genomics 10:156.

Nardi F, Carapelli A, Dallai R, Frati F. 2003. The mitochondrial genome of the olive fly *Bactrocera oleae*: Two haplotypes from distant geographical locations. Insect Mol Biol 12:605–611.

Negrisolo E, Minelli A, Valle G. 2004. Extensive gene order rearrangement in the mitochondrial genome of the centipede *Scutigera coleoptrata*. J Mol Evol 58:413–423.

Nixon KC. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15(4):407–414.

O'Clary D, Wolstenholme DR. 1983. Genes for cytochrome-c oxidase subunit-I, Urf2, and 3 transfer-RNAs in *Drosophila* mitochondrial-DNA. Nucleic Acids Res 11:6859–6872.

Ojala D, Montoya J, Attardi G. 1981. Transfer-RNA punctuation model of RNA processing in human mitochondria. Nature 290: 470–474.

Pesole G, Gissi C, De Chirico A, Saccone C. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol 48:427–434.

Podsiadlowski L, Braband A. 2006. The complete mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida). BMC Genomics 7:284.

Randall SK, Eritja R, Kaplan BE, Petruska J, Goodman MF. 1987. Nucleotide insertion kinetics opposite abasic lesions in DNA. J Biol Chem 262:6864–6870.

Roberti M, Polosa PL, Bruni F, Musicco C, Gadaleta MN, Cantatore P. 2003. DmTTF, a novel mitochondrial transcription terminating factor that recognises two sequences of *Drosophila melanogaster* mitochondrial DNA. Nucleic Acids Res 31: 1597–1604.

Satta Y, Ishiwa H, Chigusa SI. 1987. Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol Biol Evol 4:638–650.

Schneider A, Ebert D. 2004. Covariation of mitochondrial genome size with gene lengths: Evidence for gene length reduction during mitochondrial evolution. J Mol Evol 59:90–96.

Schultz BE, Chan SI. 2001. Structures and proton-pumping strategies of mitochondrial respiratory enzymes. Annu Rev Biophys Biomol Struct 30:23–65.

Sheffield NC, Song H, Cameron L, Whiting MF. 2008. A comparative analysis of mitochondrial genomes in Coleoptera (Arthropoda: Insecta) and genome descriptions of six new beetles. Mol Biol Evol 25:2499–2509.

Shi HW, Ding FM, Huang Y. 2008. Complete sequencing and analysis of mtDNA in *Phlaeoba albonema*. Chin J Biochem Mol Biol 7:604–611.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618.

Stewart JB, Beckenbach AT. 2009. Characterization of mature mitochondrial transcripts in Drosophila, and the implications for the tRNA punctuation model in arthropods. Gene 445:49–57.

Stothard P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102–1104.

Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21:537–539.

Taanman JW. 1999. The mitochondrial genome: Structure, transcription, translation and replication. Biochim Biophys Acta Bioenerg 1410:103–123.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255.

Ye W, Dang JP, Xie LD, Huang Y. 2008. Complete mitochondrial genome of *Teleogryllus emma* (Orthoptera: Gryllidae) with a new gene order in Orthoptera. Zool Res 29(3):236–244.

Yukuhiro K, Sezutsu H, Itoh M, Shimizu K, Banno Y. 2002. Significant levels of sequence divergence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkmoth, *Bombyx mandarina*, and its close relative, the domesticated silkmoth, *Bombyx mori*. Mol Biol Evol 19:1385–1389.

Zhang C, Huang Y. 2008. Complete mitochondrial genome of Oxya chinensis (Orthoptera, Acridoidea). Acta Biochimica et Biophysica Sinica 40:7–18.

Zhou Z, Huang Y, Shi F. 2007. The mitochondrial genome of Ruspolia dubia (Orthoptera: Conocephalidae) contains a short A + T rich region of 70 bp in length. Genome 50:855–866.

Zhou Z, Shi F, Huang Y. 2008. The complete mitogenome of the Chinese bush cricket, Gampsocleis gratiosa (Orthoptera: Tettigonioidea). Journal of Genetics and Genomics 35: 341–348.

Zhou Z, Huang Y, Shi F, Ye H. 2009. The complete mitochondrial genome of Deracantha onos (Orthoptera: Bradyporidae). Molecular Biology Reports 36:7–12.

Appendix 1. Orthoptera-specific primers

Name	Region	Strand	Sequence	Melting temp (°C)
OR-COX1J-3	*cox1*	J	AGGACTAGCAGGAATACCACGACG	60.0
OR-COX1N-1	*cox1*	N	TCCTACTATTCCTGCTCATGCTCC	57.9
OR-ATP6J-2	*atp6*	J	AGTCCCACAAGGAACACCACCTGC	63.3
OR-ND5J-2	*nad5*	J	AGTAGGAGCAGCCATAGCAGCAGG	62.6
OR-ND5N	*nad5*	N	TGCTGCTATGGCTGCTCCTACTCC	62.6
CA-CYTBJ	*cytb*	J	CCGAACATTACACGCAAATGGAGC	58.9
CA-CYTBN	*cytb*	N	GCTCCATTTGCGTGTAATGTTCGG	58.9
OR-ND1J	*nad1*	J	ACGAAAACGAGGTAAAGTCCCACG	59.2
OR-ND1N	*nad1*	N	TCGTGGGACTTTACCTCGTTTTCG	59.0
OR-16SJ	*16S*	J	AGAAACCGACCTGGCTCACGCCGG	67.3
OR-16SN	*16S*	N	TCAGACCGGCGTGAGCCAGGTCGG	68.7
OR-12SJ	*12S*	J	CGTATAACCGCGGCTGCTGGCACG	66.3
OR-12SN	*12S*	N	CGTGCCAGCAGCCGCGGTTATACG	66.3