Progress in the Researches on Insect Mitochondrial Genome and Analysis of Gene Order

Li Hu  Gao Jianyu  Liu Haiyu  Cai Wanzhi*

(Department of Entomology, China Agricultural University, Yuanmingyuan West Road, Beijing 100193, China; *corresponding author, email: caiwz@cau.edu.cn)

Insect mitochondrial genome is a double-stranded circular genomes which range from 14, 503 bp to 19, 517 bp in size. Nearly all the sequenced insect mitochondrial genomes encode 37 genes: two for rRNAs, 13 for proteins and 22 for tRNAs. This review compares and summarizes the features of complete mitochondrial genomes from 175 sequenced species of insects in 22 orders. The genomic organization, contents, gene order, and rearrangements of gene order are analyzed.

Key words  insect; mitochondrial genome; gene order; phylogeny

Mitochondrial (mt) genomes have become a major data sources for comparative genomics and play an important role in metabolism[1], apoptosis[2], disease[3] and aging[4]. In insects, mtDNA is typically a small genome, less than 16 kb, in combination with a generally conserved gene completement and rapid rate of nucleotide substitution, provide an ideal system for a wide range of comparative studies that have furthered the understanding of genome evolution.

Due to the relative ease of amplification and sequencing, in the last two decades the insect mitochondrial genome has become the most commonly used molecular marker for population genetics, phylogeography, molecular diagnostics and phylogenetic studies. In the present review, the basic features of complete mitochondrial genomes from 175 species of insects in 22 orders are surveyed and summarized with respects of their genomic organization, contents, and gene order.

1 Structure and composition of mitochondrial genomes of insects

To date the complete or near complete mitochondrial genomes have been sequenced from 175 insect species in 22 orders. Most species come from the order Diptera, Hemiptera, Hymenoptera, Lepidoptera and Orthoptera.

The mitochondria of insects contain their own double-stranded circular genomes which range from 14, 503 bp[5] to 19, 517 bp in size[6]. Despite the extraordinary range in size, the gene content of the molecule is remarkably conserved. With few exceptions it encodes 37 genes including 13 protein-coding subunits (COI-III, CytB, ND1-6, ND4L, ATP6 and ATP8) from three of the oxidative phosphorylation complexes, plus the 2 ribosomal RNA (rrnS and rrnL) and 22 transfer RNA (trnI, trnQ, trnM, trnW, trnC, trnY, trnL1, trnK, trnD, trnG, trnA, trnR, trnN, trnS1, trnE, trnF, trnH, trnT, trnP, trnS2, trnL2, trnV) genes necessary to translate the protein-coding genes[7]. In addition, mt genomes of insects contain one large AT-rich non-coding control elements, which are involved in the initiation and regulation of mt transcription and replication, and are therefore referred to as the AT-rich region or the mitochondrial control region (CR), which is usually present between srRNA and tRNA-I. The use of the term “AT-rich region” to describe the control region should be reconsidered even though each control region has one or several more AT-rich subregions, because this region is not always the most AT-rich part of the mt genome[8].

Except for the control region, the entire genome comprises compactly arranged coding sequences with no introns and with very few intergenic nucleotides, and neighboring genes may even shortly overlap in some cases. For instance, besides the control
region there are two major non-coding regions of 232 and 273 bp in length that are located at the gene junction tRNA Q-E and tRNA W-S2 in Bothriometopus macrocnemis[9].

Overlapping genes found among many mtDNAs leads to speculation that the “polycistron model” may not universally apply, since it would not be possible to release full-length RNAs of each overlapping message from the same transcript[10].

Both the largest and smallest mitochondrial genome comes from the order Diptera. The former is Drosophila melanogaster, which is 19, 517 bp in size[6, 11], and the latter is Rhapalomyia pomum, which is 14, 503 bp in size[5]. The reason for the difference in size of the mitochondrial genome is the different length of the control region. Generally the size of the control region of mtDNA range from 70 bp (Orthoptera: Ruspolia dubia)[22] to 4.6 kb (Diptera: Drosophila melanogaster)[6, 11]. The views that mtDNAs are selected for extreme economy of size, while apparent in some cases, is refuted by mitochondrial genomes with very large amounts of non-coding sequence, as has been found in some species.

The genome which contains the standard set of 22 tRNA genes usually presents in insects mtDNAs, but there are some exceptions (Table 1). For example, the genome of Bothriometopus macrocnemis encodes 25 tRNA genes, 21 of the 22 tRNA genes common to most mt genomes of insects plus three additional copies of the tRNA-Y, and an additional copy of the tRNA-V. The gene for the tRNA-A is lacking. In addition, remnants of two additional copies of tRNA are present. The two tRNA-like pseudogenes were found between tRNA-N and tRNA-T and between tRNA-S2 and COI. In both regions a cruciform secondary structure resembling a tRNA could be found. However, neither could form a functional anticodon loop[9].

The total AT content calculated over the whole molecule ranges from 63%[13] to 88%[14]. This conforms to the well-know A+T bias of insect mitochondrial genomes. In some mtDNA all genes are transcribed from one strand such as the screamer louse Bothriometopus macrocnemis[9], whereas in others the genes are distributed between the two strands. In most cases, the majority of genes are located on the plus or J-strand, the remainder having opposite polarity and being oriented on the minus or N-strand. Except for different amount of genes on the two strands, there is an asymmetrical compositional bias across the two strands. In an olive fly Bactrocera oleae[15], protein coding genes on the J-strand show a fairly similar A% and T%, whereas those encoded on the N-strand have a much higher proportion of T than A. The presence of such a bias suggests that genes may be subject to additional substitutional pressures (directional selection) following a translocation that involves a change in the coding strand. This phenomenon may have important consequences on molecular phylogenetic studies when single mitochondrial genes are used to compare distantly related taxa, in which the same gene product may be encoded on different strands[15].

Canonical initiation codons (ATA or ATG), encoding the amino acid methionine, are usually used in most PCGs[16]. However, there are apparent exceptions to the ATN rule[15]. Some genes start with non-standard codons as it often happens in mtDNAs of insects. And the translation initiation signal for the COI gene is a good case in point. In Drosophila melanogaster and D. yakuba[17], no ATN in-frame codon is found in the proximity of the presumed start site of the gene. Although a TCG triplet opens the COI gene reading frame, a tetranucleotide ATAA, positioned directly upstream was proposed as an initiator[19]. Two mechanisms that would permit translation to start at this sequence were suggested that either initiation occurs at the ATA codon following a +1 frame shift of the translation machinery to restore the reading frame and resume a correct translation, or that the anticodon of the initiation N-formylmethionine tRNA might permit the ATAA sequence to be recognized as a single codon.

In most cases PCGs terminate with the complete termination codon TAA or TAG, but in all other cases, stop codons are truncated (T or TA) and their functionality probably recovered after a post-transcriptional polyadenylation[20]. These abbreviated stop codons are found in PCGs that are followed by a downstream tRNA gene, suggesting that the secondary structure information of the tRNA genes could be responsible for the correct cleavage of the polycistronic transcript[21]. However, there are also direct junctions between two PCGs (ATP8- ATP6, ATP6-COIII, ND6-CytB and ND4L-ND4) where other cleavage signals, different from tRNA gene secondary structures, may be involved in the processing of the polycistronic primary transcript[22]. In this respect, it is found that possibly stem-loop structures in the 3’ end of those protein-coding genes which are not flanked by tRNAs[23-25].

2 Gene order and rearrangement of mt genomes
Although animal mitochondrial sequences are known to evolve rapidly, arrangements of insect mitochondrial genes are relatively stable, with some having been conserved over long periods of evolutionary time. Because of the low rearrangement rate as well as several other characteristics, mitochondrial gene arrangements promise to be a useful data set for the study of deep divergences of insects. Indeed, several problematic relationships have already been convincingly resolved using this data set[26, 27].

Whether certain types of gene rearrangement are more common than others is a question. Genome rearrangements can be characterized in several aspects: (1) the types of genes rearranged, tRNAs only (termed minor rearrangements) or also protein-coding and rRNA genes (major rearrangements); (2) whether genes are translocated along the same strand or inverted between coding strands; and (3) the localization of the rearrangement (between local or distant gene blocks)[28]. A given gene rearrangement may exhibit one or more of these aspects. In addition, there are different mechanistic models of how genome arrangements occur.

The best supported model attributes gene rearrangements to the partial duplication of mtDNA caused by errors in replication, such as erroneous initiation or termination, or strand slippage and mispairing, followed by the loss of one copy of each duplicated gene. It is commonly assumed that the loss of one of the two copies of each duplicated gene happens at random, and the model described is referred to as the duplication-random loss model[29-31].

Another most commonly mechanism invoked is the model: mtDNA duplication and nonrandom gene loss. According to this model, the destiny of each gene copy in the duplicated region is predeterminated by its transcriptional polarity and location in the genome[32].

Because translocations could not be explained by the alternative replication-slipage-based models, the incidence of genome translocations has been used as evidence for the occurrence of intra-mitochondrial recombination[33].

As additional mt genomes have been sequenced, however, the phylogenetic utility of gene rearrangement in understanding insect evolution now appears less than previously thought because most insect mt genomes have retained the ancestral gene order and there is not enough phylogenetic signal to resolve interordinal or deeper relationships from the existing gene rearrangements.

In Insecta, mitochondrial gene rearrangements have been found in 83 species belonging to 11 orders, with the others exhibiting the same gene order observed in Drosophila yakuba[18]. Rearrangements appear much common in tRNA genes and very rare in protein-coding genes and rRNA genes. Rearrangements of protein-coding genes were found in three orders: three species in Hemiptera[34, 37], one species in Lepidoptera[36], one species in Pscoptera[37] and one species in Coleoptera. Both rearrangements of protein-coding genes and rRNA genes were observed in the orders Thysanoptera[39] and Phthiraptera[9, 38].

Variability in gene order can be observed within groups at many taxonomic levels, and it has been found that four hemipteroid orders (Hemiptera, Thysanoptera, Pscoptera, and Phthiraptera) and Hymenoptera have increased rates of gene rearrangements in the mt genomes.

In the order Thysanoptera, rearrangements of 24 genes appear in the species Thrips imagines[9], and translocations are found in eight of them. This species has two very similar non-coding regions, with 18 genes rearranged between them.

Almost all 37 genes are transposed from their ancestral position in the two sequenced phthirapteran species. So lice are an ideal group in which to study rearrangements due to the heightened rearrangement rate. It is found that 11 of the 13 protein-coding genes have 3’ end stem-loop structures which may allow mRNA processing without flanking tRNAs in bothriometopus macrocnenis and so facilitate gene rearrangements.

Within Hemiptera, there are available a number of completely sequenced mitochondrial genomes that have variations in the gene order[8, 39]. All observed rearrangements fall into three types: A: translocations and transpositions of tRNA genes on the basis of the ancestral gene order; B: L-Q-M-ND2-W-C-Y block transposed to the end of the control region, with the translocation of some tRNA genes; C: transpositions of DNA fragment containing COIII-G-ND3-A-R-N on the basis of type B. In type C, there is variation in the mitochondrial position into which these genes are transposed. In addition, there are differences with respect to the retention of the number and the order of the excised tRNA genes at the mitochondrial location in which the genes are inserted. The maximal insertion involves all of the genes from the excised fragment in their original order A-R-N-ND3-G-COIII, the minimal insertion involves ND3-G-COIII. In all insertions, the transcription direction is altered from that in the original position[34].

Surprisingly, among ten completely sequenced mitochondrial genomes in the order Hymenoptera, most of tRNA genes are...
transposed\textsuperscript{[9, 14, 40-45]}. The new orders of these genomes seem haphazard.

By contrast, rearrangements of mt genomes from many other insect orders appear stable in species and gene orders are similar to that of proposed ancestor.

Most orthopterans just contain a tRNA gene rearrangement from the insect ground plan $K-D$ swapping to $D-K$\textsuperscript{[46-52]}, except the species \textit{Troglophilus neglectus} with an additional $tRNA-L$\textsuperscript{[46]} and \textit{Teleogryllus emma} with the translocation of $tRNA-N$ and $tRNA-E$\textsuperscript{[53]}. Three species in Neuroptera only have the translocation of $tRNA-W$ and $tRNA-C$\textsuperscript{[54, 55]}. In Coleoptera, the region of the mt genome consisting of genes encoding for $I-Q-M$\textsuperscript{[56, 57]} and $I-Q-M-ND2-W-C-Y$ can be transposed from its ancestral position to the end of the control region. The translocation of $tRNA-I-Q$ and $tRNA-M$ is the common rearrangements of mt genomes in Lepidoptera\textsuperscript{[36, 58-67]}. Changes in the Ephemeroptera, Phasmatoidea, Archaeognatha and Diptera are the gene arrangement in the location of $I-Q-M$, $W-C-Y$ and $A-R-N-S1-E-F$ clusters\textsuperscript{[35, 68-73]}.

It seems that the most common changes in mt genomes are of genes near the large non-coding region or of the tRNAs of regions: $I-Q-M$, $W-C-Y$, $K-D$ and $A-R-N-S1-E-F$, especially the region $A-R-N-S1-E-F$. This points to a role for the origins of replication in gene order translocation and, perhaps, suggests a closer look be taken for a second-strand origin in arthropod mtDNA near the $A-R-N-S1-E-F$ region\textsuperscript{[27]}.

Rearrangements appear much more common in insects and variability in gene order can be found within groups at many taxonomic levels. With few exceptions, gene arrangements are relatively stable within major groups, but variable between them, and comparisons of these gene arrangements have great potential for resolving some of the deepest branches of insect phylogeny.

3 Conclusions

Complete mitochondrial genome sequences are being determined at an ever-increasing rate, providing many new opportunities for comparing genome structure. MtDNA data have played a major role in studies of molecular evolution that have helped to refine models of evolution and understand how rates and patterns of substitution vary across sequences and over time. With the completion of whole genome sequencing projects accelerating, this will serve as a model for interpreting broader aspects of genome evolution and insect phylogeny. In addition to expanding this data set with broader taxonomic representation, it is important that studies address the molecular mechanism leading to rearrangement and develop better theoretical models and computer reconstructions.

Acknowledgements

This research is supported by grants from the National Natural Science Foundation of China (Nos. 30825006, 30970394) and the Doctoral Program of Higher Education of China (No. 200800190015).

References

[9] Cameron SL, Johnson KP, Whiting MF. The mitochondrial genome of the screamer louse \textit{Bothriometopus} (Phthiraptera: Ischnocera): effects of


[55] Cameron SL, Sullivan J, Song H, Miller KB, Whiting MF. A mitochondrial genome phylogeny of the Neuroptera (lace-wings, alderflies and


[64] Cameron SL, Whiting MF. 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. 2008, 408(1-2): 112-123.


