Insect Biochemistry and Molecular Biology 113 (2019) 103208

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Description of soybean aphid (*Aphis glycines* Matsumura) mitochondrial genome and comparative mitogenomics of Aphididae (Hemiptera: Sternorrhyncha)

Hojun Song[a,*], Ravi Kiran Donthu[b,c], Richard Hall[d], Lawrence Hon[e], Everett Weber[b,c], Jonathan H. Badger[f], Rosanna Giordano[b,c]

[a] Department of Entomology, Texas A&M University, College Station, TX, USA
[b] Puerto Rico, Science, Technology & Research Trust, San Juan, PR, USA
[c] Know Your Bee, Inc, San Juan, PR, USA
[d] Pacific Biosciences, Menlo Park, CA, USA
[e] Color Genomics, Burlingame, CA, USA
[f] Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institute of Health, DHHS, Bethesda, MD, USA

ARTICLE INFO

Keywords:
Aphididae
Comparative analysis
Repeat region
Phylogeny

ABSTRACT

The complete mitochondrial genome of the soybean aphid (*Aphis glycines* Matsumura), a major agricultural pest in the world, is described for the first time, which consists of 13 protein-coding genes, 22 tRNA genes, 2 rRNA genes, as well as a large repeat region between tRNA-Glu and tRNA-Phe, and an AT-rich control region. The 17,954 bp mtgenome is the largest reported from the family Aphididae, and its gene order follows the ancestral insect mtgenome except for the repeat region, which contains a 195 bp unit repeated 11.9 times, representing the highest reported repeats among the known aphid mtgenomes to date. A new molecular phylogeny of Aphidae is reconstructed based on all available aphid mtgenomes, and it is shown that the mtgenome data can robustly resolve relationships at the subfamily level, but do not have sufficient phylogenetic information to resolve deep relationships. A phylogeny-based comparative analysis of mtgenomes has been performed to investigate the evolution of the repeat region between tRNA-Glu and tRNA-Phe. So far, among aphids, 13 species are known to have this repeat region of variable lengths, and a phylogenetic analysis of the repeat region shows that a large proportion of the sequences are conserved across the phylogeny, suggesting that the repeat region evolved in the most recent common ancestor of Aphidinae and Eriosomatinae, and that it has gone through numerous episodes of lineage-specific losses and expansions. Combined together, this study provides novel insights into how the repeat regions have evolved within aphids.

1. Introduction

The soybean aphid (*Aphis glycines* Matsumura) (Hemiptera: Sternorryhncha: Aphididae) is one of the most important agricultural pests of soybean in the world. Originally from Asia, the soybean aphid has recently invaded the New World, first discovered in Wisconsin in 2000 (Ragsdale et al., 2004), and now it is firmly established in the Midwestern and Eastern United States as well as Canada. The infestation of the soybean aphid can cause shortened plant height and sooty mold growth resulting in significant crop loss (Kim et al., 2008). The soybean aphid belongs to the genus *Aphis* Linnaeus, 1758, which contains at least 600 species distributed worldwide, several of which are important agricultural pests including the cotton aphid (*A. gossypii* Glover), the apple aphid (*A. pomi* de Geer), the cowpea aphid (*A. craccivora* Koch), and the black bean aphid (*A. fabae* Scopoli). Within the family Aphididae (Hemiptera: Sternorryhncha), which include small phloem-feeding insects, many of which are serious plant pests and vectors of plant viruses (Dixon, 1997; Ragsdale et al., 2011), there are currently more than 5100 species of aphids known belonging to 23 subfamilies (Favret, 2018). The soybean aphid belongs to the subfamily Aphidinae, which is the most diverse subfamily, containing more than 3000 species.

Despite the agricultural importance of the soybean aphid, its mitochondrial genome is currently incompletely known. Mitochondrial genomes (mtgenomes) are one of the smallest organellar genomes (Boore, 1999) composed of circular, double-stranded DNA, which is

* Corresponding author.
E-mail address: hsong@tamu.edu (H. Song).

https://doi.org/10.1016/j.ibmb.2019.103208
Received 26 January 2019; Received in revised form 26 June 2019; Accepted 25 July 2019
Available online 15 August 2019
0965-1748/

Table 1
Mitochondrial genome description of soybean aphid.

Gene	Anticodon	Start Codon	Stop Codon	Strand	Position	Intergenic Space
tRNA-Ile	GAU			+	1–64	−3
tRNA-Gln	UUG			–	62–127	4
tRNA-Met	CAU			+	132–197	0
ND2		ATT	T	+	198–1173	0
tRNA-Trp	UCA			+	1174–1235	−8
tRNA-Cys	GCA			–	1228–1293	1
tRNA-Tyr	GUA			–	1295–1359	1
COI		ATA	T	+	1361–2891	0
tRNA-Leu2	UAA			+	2892–2959	3
COII		ATA	TAA	+	2963–3634	2
tRNA-Lys	CUU			+	3637–3709	0
tRNA-Asp	GUC			+	3710–3772	0
ATP8		ATT	TAA	+	3773–3931	1
ATP6		ATT	TAA	+	3933–4565	−1
COIII		ATG	TA	+	4565–5349	0
tRNA-Gly	UCC			+	5350–5412	0
ND3		ATT	TAA	+	5413–5766	0
tRNA-Ala	UGC			+	5767–5830	−1
tRNA-Arg	UCG			+	5830–5893	−1
tRNA-Asn	GUU			+	5893–5958	−1
tRNA-Ser1	GCU			+	5958–6021	2
tRNA-Glu	UUC			+	6024–6086	0
repeat region				NA	6087–8408	0
tRNA-Phe	GAA			–	8409–8474	0
ND5		ATT	TAA	–	8475–10145	57
tRNA-His	GUG			–	10203–10266	0
ND4		ATA	T	–	10267–11584	−1
ND4L		ATA	TAA	–	11584–11874	1
tRNA-Thr	UGU			+	11876–11938	3
tRNA-Pro	UGG			–	11942–12009	1
ND6		ATT	TAA	+	12011–12502	3
CYTB		ATG	TAA	+	12506–13621	6
tRNA-Ser2	UGA			+	13628–13693	10
ND6		ATT	TAA	–	13704–14639	0
tRNA-Leu1	UAG			–	14640–14704	0
16S				–	14705–15965	0
tRNA-Val	UAC			–	15966–16027	0
12S				–	16028–16796	0
control region				NA	16797–17954	0

about 14,000 to 18,000 bp in length, and usually encode 37 genes (13 protein coding, 22 transfer RNA, and 2 ribosomal RNA genes) and a non-coding control region of variable length (Boore, 1999; Taanman, 1999; Wolstenhome, 1992). For Aphididae, there are currently 26 complete and 22 partial mtgenomes available in GenBank. Since the description of the first aphid mtgenome (*Schizaphis graminum*) by Thao et al. (2004), many aphid mtgenomes have been published (Li et al., 2017; Ren et al., 2016; Song et al., 2016a, 2016b; Sun et al., 2017; Wang et al., 2013, 2015, 2014; Zhang et al., 2013, 2016a), which now cover most of the phylogenetic diversity within the family, including 15 subfamilies. Several comparative analyses of aphid mtgenomes have suggested that they have retained the ancestral insect mtgenome features with the exception of what appears to be a phylogenetically conserved insertion of non-coding repeat region of variable length between tRNA-Glu and tRNA-Phe found in the subfamily Aphidinae (Wang et al., 2013, 2015). However, there has not been a recent synthesis of the comparative mitogenomics within Aphididae. Until now, only two *Aphis* mtgenomes have been completely sequenced: *A. gossypii* and *A. craccivora* (Song et al., 2016b; Sun et al., 2017; Zhang et al., 2016b). Wang et al. (2013) sequenced five aphid mtgenomes, and included the soybean aphid, but they were unable to sequence through the repeat region, resulting in a partial mtgenome for this species (13,002 bp). Therefore, we currently do not know the characteristic of the repeat region in the soybean aphid and how it compares to other aphid species.

Mitochondrial genome data have been frequently used for insect phylogenetics (Cameron et al., 2004, 2006, 2007; Chen et al., 2017; Fenn et al., 2008) because they are considered as a versatile marker that can broadly resolve relationships at different phylogenetic scales. Some insect groups have unique mtgenomes features, such as gene rearrangements, which could be considered molecular synapomorphies for certain clades (reviewed in Cameron et al., 2014). For Aphididae, there are two recent phylogenetic analyses using mtgenome data that aim to resolve higher-level relationships within the family (Chen et al., 2017; Ren et al., 2017). Chen et al. (2017) included 35 aphid species and Ren et al. (2017) included 37 aphid species, both of which used the same 19 aphid mtgenomes available from GenBank. Although both studies were done rigorously, the resulting phylogenies differ greatly. For example, although both studies found the subfamily Aphidinae, to which the soybean aphid belongs, to be monophyletic, the internal relationship within the subfamily differed, and Chen et al. (2017) found Aphidinae to be sister to a clade consisting of Phyllaphidinae, Calaphidinae, and Saltusaphidinae, while Ren et al. (2017) found it to be sister to a clade consisting of Phloemyzinae and Thelaxinae. The differences between the two studies are likely due to differences in taxon sampling. To resolve this conflict, it is important to simultaneously analyze all the available mtgenome data.

In this study, we describe the complete mtgenome of the soybean aphid (Biotype 1) for the first time and compare and contrast this mtgenome with other known aphid mtgenomes. We also reconstruct the most extensive phylogeny of Aphididae based on all available mtgenome sequences. Due to the increase in taxon sampling contained herein, which have not been previously combined and examined, we recovered novel relationships within the Aphididae. We also examine the evolution of the repeat region between tRNA-Glu and tRNA-Phe using a character mapping approach as well as the phylogenetic

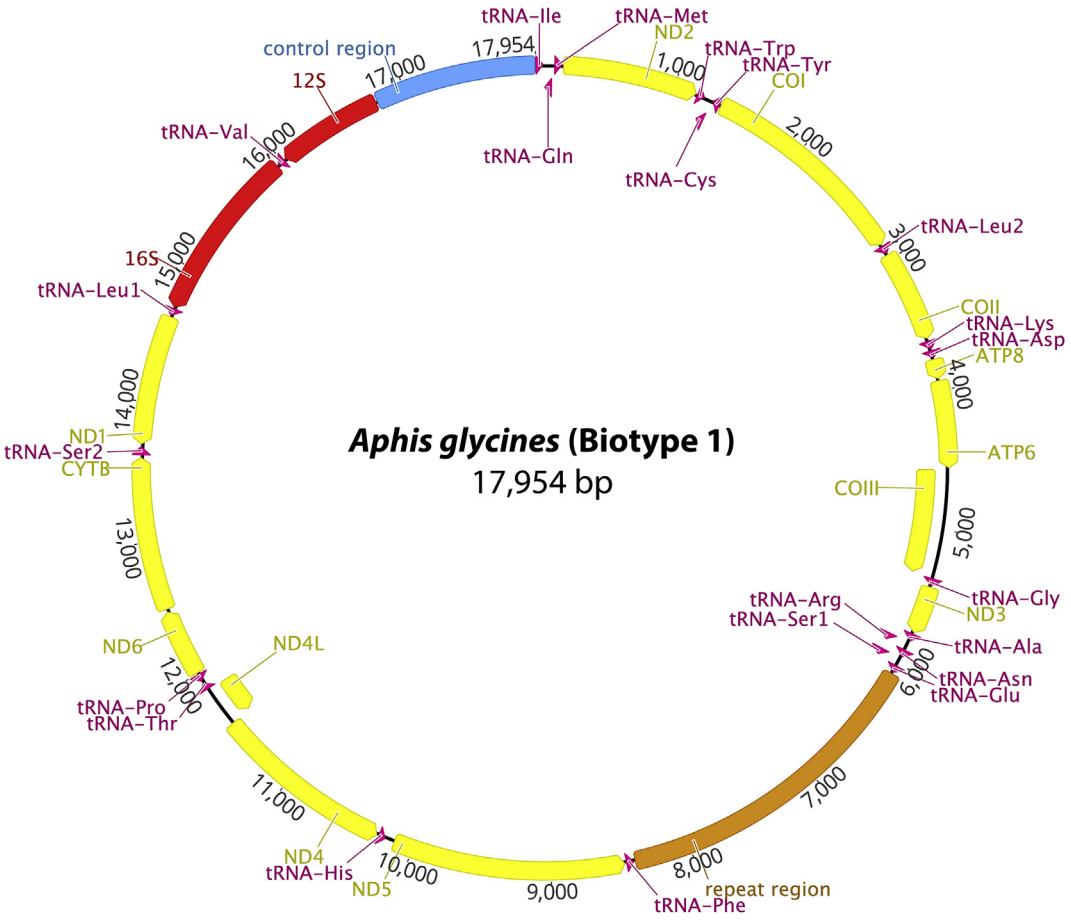

Fig. 1. Circularized mtgenome of the soybean aphid.

analysis of the repeat region sequences of the 13 species known to have this trait. We show that a large proportion of the repeat region sequences are conserved across the phylogeny, but there have been numerous episodes of lineage-specific losses and expansions of the repeat regions throughout the diversification of aphids.

2. Materials and methods

2.1. Data generation

The generation of the mtgenome sequence data was part of a larger genome sequencing project for the soybean aphid. The genome of the soybean aphid was sequenced using a combination of PacBio, Illumina and 454 reads. All sequencing reads from the three technologies were assembled using the MaSuRCA assembler (Zimin et al., 2013). For a detailed description of the assembly and annotation of the soybean aphid genome, please see Giordano et al. (in press).

The mitochondrial sequences of the soybean aphid were extracted in the following manner. Raw reads were first aligned to the assembled contigs using BLASR (Chaisson and Tesler, 2012), and contigs with higher than expected mean coverage were taken and used in a BLAST search against NCBI "Nucleotide collection (nr/nt)" via the NCBI web interface to identify mitochondrial contigs.

To confirm that the long repeat region between tRNA-Glu and tRNA-Phe was real and not an assembly error, two fasta files were created from the initial assembly, one spanning from tRNA-Ile to tRNA-Phe and another spanning from tRNA-Glu to the control region, to be used as reference sequences to extract mitochondrial sequences from raw Illumina data. We used 'Map to Reference' tool in Geneious 10.0.9 (Biomatters) and selected the Geneious mapper with medium to low sensitivity to search for short reads that mapped to the reference sequences three times. The resulting consensus sequences were inspected to confirm the nucleotide sequences for the long repeat region.

2.2. Annotation

The assembled mtgenome fasta file was uploaded to MITOS (Bernt et al., 2013) to identify open reading frames (ORFs) and tRNAs. The initial MITOS annotation was used as a guideline to delimit gene boundaries, and start and stop codons of each protein-coding gene (PCG) were manually identified in Geneious, following the recommendation by Cameron (2014a). The secondary structures of tRNAs were predicted using MITOS. The resulting annotated mtgenome was circularized in Geneious and submitted to GenBank, with accession number MK111111.

2.3. Phylogenetics

We sampled a total of 51 taxa, representing all available aphid mtgenomes, including the soybean aphid generated in this study. For the remaining 50 taxa, we obtained the mtgenome sequences from GenBank. We included two non-aphid outgroup taxa, *Adelges laricis* (Adelgidae) and *Daktulosphaira vitifoliae* (Phylloxeridae) based on previous phylogenetic studies which used these two taxa as outgroups (Chen et al., 2017; Ren et al., 2017), and used Adelgidae as the root taxon. Thus, we included 49 aphid taxa for our ingroup sampling. The taxonomic classification associated with the GenBank data was often outdated and incorrect, thus we used the Aphid Species File (Favret, 2018) to update the taxonomy for all included taxa. Of the 51 taxa, 27 taxa had complete mtgenomes, 5 taxa had partial mtgenomes, and 19

Table 2
Comparison of base composition (AT%) across the mtgenomes of Aphididae. ATP refers to the mean AT% ATP6 and ATP8; COX refers to COI–COIII, and CYTB; NADH refers to ND1-ND6, rRNA refers to 16S and 12S; tRNA refers to 22 tRNA genes; CR refers to control region; Repeat refers to the repeat region between tRNA-Glu and tRNA-Phe.

GenBank #	Family	Subfamily	Tribe	Species	Sequence length	Mtgenome	ATP	COX	NADH	rRNA	tRNA	CR	Repeat	Repeat region length
KP722589	Adelgidae			*Adelges laricis*	12,146	83.0	85.9	79.9	85.6	–	–	–	–	unknown
DQ021446	Phylloxeridae	Phylloxerinae	Phylloxerini	*Daktulosphaira vitifoliae*	12,349	83.7	87.1	80.7	86.3	85.1	85.2	85.9	–	absent
KP722590	Aphididae	Aiceoninae		*Aiceona himalaica*	9991	82.6	88.2	80.4	85.9	–	–	–	–	unknown
KP722588		Anoeciinae		*Anoecia fulviabdominalis*	11,933	81.0	86.7	77.6	85.3	–	–	–	–	unknown
KX447142		Aphidinae	Aphidini	*Aphis craccivora (New World)*	15,305	83.7	85.9	79.9	86.1	85.2	85.5	82.8	85.4	178
NC_031387				*Aphis craccivora (Old World)*	15,308	83.7	86.3	79.8	86.0	85.2	85.7	82.8	86.0	179
KT889380				*Aphis craccivora (Zhengzhou)*	14,425	83.3	85.2	79.9	85.9	84.4	85.2	–	–	unknown
This Study				*Aphis glycines (Biotype 1)*	17,954	83.5	83.1	79.6	86.4	84.8	85.4	84.2	89.8	2322
KC840675				*Aphis glycines (China)*	13,002	83.1	83.5	79.6	86.9	84.8	84.3	83.5	–	unknown
NC_024581				*Aphis gossypii*	15,869	83.4	85.2	79.4	86.4	84.6	85.6	84.7	88.3	784
KT447631				*Rhopalosiphum padi*	15,205	83.1	84.3	78.8	86.1	84.6	85.3	87.1	–	unknown
NC_006158				*Schizaphis graminum*	15,721	83.7	84.6	79.5	86.7	84.8	85.6	86.8	84.2	608
NC_011594			Macrosiphini	*Acyrthosiphon pisum*	16,971	82.8	87.0	80.1	87.1	84.5	85.3	93.3	84.7	1488
NC_022682				*Cavariella salicicola*	16,317	83.4	87.7	79.7	85.9	84.3	85.2	85.5	91.5	702
NC_022727				*Diuraphis noxia*	15,784	84.3	87.4	80.7	87.3	84.9	85.4	86.6	90.7	644
NC_029727				*Myzus persicae*	17,382	83.4	86.9	79.6	85.9	84.8	85.8	87.1	91.2	307
KC840676				*Pterocomma pilosum*	12,529	82.3	85.5	78.7	85.6	84.4	85.8	76.3	–	unknown
NC_024683				*Sitobion avenae*	15,180	83.8	85.6	79.9	86.8	84.1	84.8	93.0	88.9	261
KP722573		Calaphidinae	Panaphidini	*Eucallipterus tiliae*	12,146	82.7	87.8	79.6	85.9	–	–	–	–	unknown
KP722568				*Takecallis arundinariae*	9455	80.8	84.6	78.0	85.0	–	–	–	–	unknown
KP722584		Chaitophorinae	Chaitophorini	*Chaitophorus saliniger*	11,158	81.1	85.0	77.1	85.0	–	–	–	–	unknown
KP722572				*Periphyllus koelreuteriae*	11,469	81.2	86.1	76.8	85.4	–	–	–	–	unknown
KP722585			Siphini	*Chaetosiphella stipae*	9564	82.0	86.8	79.0	86.7	–	–	–	–	unknown
NC_033352		Eriosomatinae	Eriosomatini	*Eriosoma lanigerum*	15,640	84.3	88.6	80.7	87.2	85.0	86.3	88.0	–	absent
NC_035314			Fordini	*Baizongia pistaciae*	15,602	84.5	88.1	81.0	87.1	84.8	86.6	89.1	89.9	395
MF043985				*Kaburagia rhusicola ovatirhusicola*	16,184	83.3	86.1	79.3	86.1	84.7	86.2	82.6	–	absent
MF043987				*Kaburagia rhusicola rhusicola*	16,159	83.1	86.2	79.1	86.0	84.8	86.2	81.7	–	absent
NC_036065				*Melaphis rhois*	15,436	82.5	85.2	77.7	85.1	84.1	86.0	84.9	–	absent
NC_035310				*Nurudea choui*	15,308	83.8	85.7	79.6	86.7	85.3	86.4	86.3	–	absent
NC_035311				*Nurudea ibofushi*	16,054	83.7	86.5	80.1	86.4	84.5	86.3	86.0	89.6	670
NC_035316				*Nurudea meitanensis*	15,301	83.8	85.7	79.6	86.8	85.3	86.5	86.1	–	absent
NC_035301				*Nurudea shiraii*	15,389	83.9	86.5	79.7	86.7	84.6	86.2	84.5	–	absent
NC_035313				*Nurudea yanoniella*	15,858	83.4	86.2	79.3	86.2	84.3	86.2	85.4	84.5	452
NC_032386				*Schlechtendalia chinensis*	16,047	83.8	86.7	80.1	86.5	84.6	86.3	85.5	89.0	670
NC_035315				*Schlechtendalia elongallis*	16,191	83.7	87.5	79.7	86.3	85.0	86.4	87.1	–	absent
NC_035312				*Schlechtendalia flavogallis*	16,150	83.2	86.2	79.2	86.1	84.8	86.3	81.9	–	absent
NC_035302				*Schlechtendalia peitan*	15,609	83.8	88.2	79.9	86.4	85.0	85.8	87.8	–	absent
NC_024926		Greenideinae	Cervaphidini	*Cervaphis quercus*	15,272	84.4	89.1	79.6	87.8	85.4	86.5	90.7	–	absent
KP722580			Greenideini	*Greenidea kuwanai*	12,146	84.6	89.6	81.7	87.3	–	–	–	–	unknown
KP722586		Hormaphidinae	Cerataphidini	*Ceratovacuna lanigera*	11,549	82.7	87.3	79.5	86.3	–	–	–	–	unknown
NC_029495			Hormaphidini	*Hormaphis betulae*	15,088	82.2	84.2	77.6	85.5	83.6	84.7	84.0	–	absent
KP722574			Nipponaphidini	*Neothoracaphis yanonis*	12,146	83.4	86.7	80.1	86.9	–	–	–	–	unknown
KP722583		Lachninae	Eulachnini	*Cinara tujafilina*	10,866	82.4	88.6	79.3	85.7	–	–	–	–	unknown
KP722566			Tuberolachnini	*Tuberolachnus salignus*	11,162	82.3	85.5	79.1	85.6	–	–	–	–	unknown
KP722577		Macropodaphidinae		*Macropodaphis* sp.	11,072	82.1	87.1	79.4	85.2	–	–	–	–	unknown
NC_033410		Mindarinae		*Mindarus keteleerifoliae*	15,199	84.2	89.3	80.2	87.2	84.9	85.8	88.9	–	absent
KP722571		Phloeomyzinae		*Phloeomyzus passerinii*	12,146	84.9	88.7	81.6	87.8	–	–	–	–	unknown
KP722570		Phyllaphidinae		*Phyllaphis fagi*	9251	82.3	86.0	78.8	87.0	–	–	–	–	unknown
KP722569		Saltusaphidinae	Saltusaphidini	*Saltusaphis sp*	8907	82.3	86.8	80.7	85.7	–	–	–	–	unknown
KP722567			Thripsaphidini	*Thripsaphis sp*	9000	81.9	86.2	79.3	86.2	–	–	–	–	unknown
KP722578		Thelaxinae	Thelaxini	*Kurisakia onigurumii*	11,166	82.9	88.2	80.3	85.0	–	–	–	–	unknown

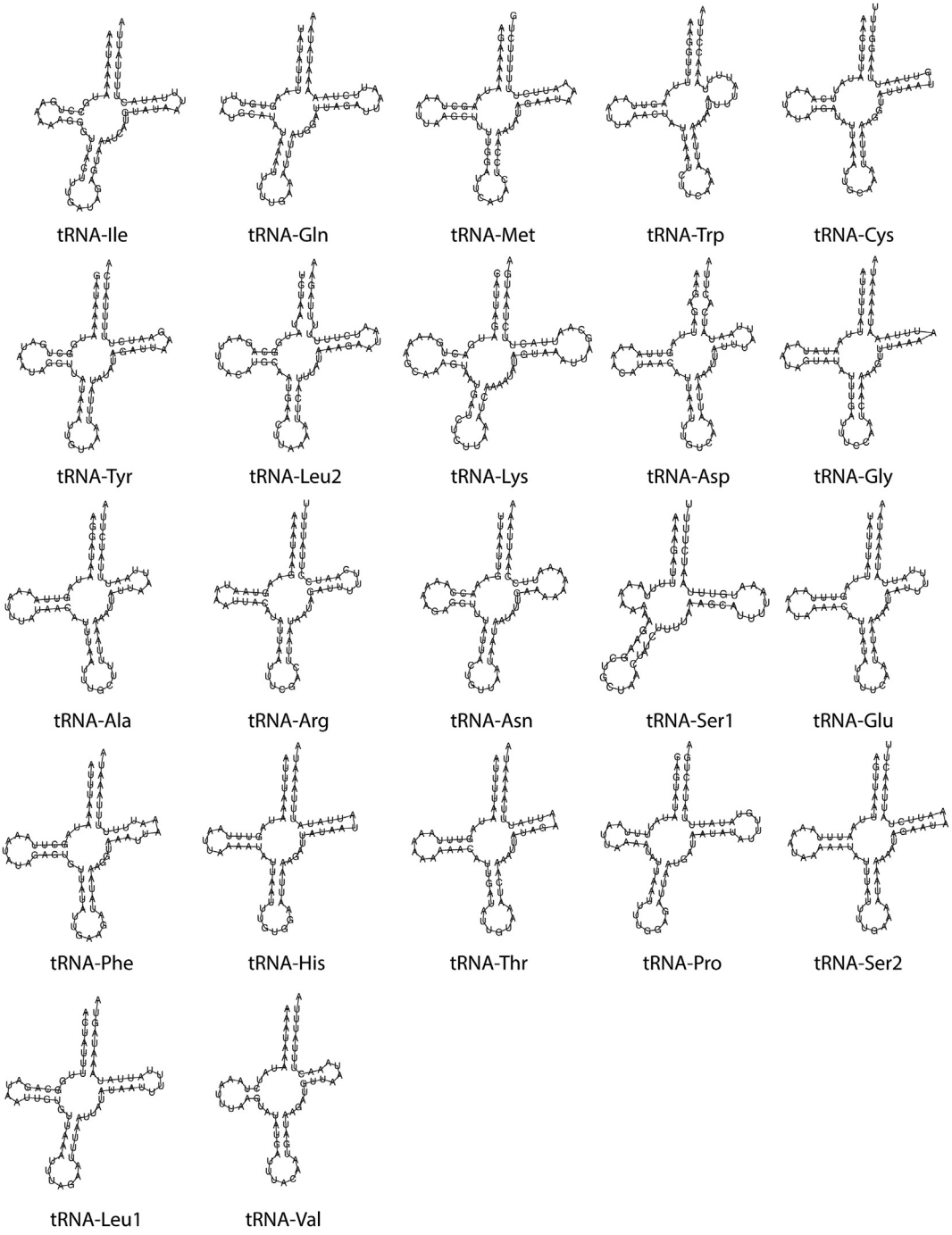

Fig. 2. Secondary structure of 22 transfer RNA genes of the soybean aphid.

taxa only had the PCGs.

For the PCGs, we aligned based on the conservation of reading frames by first translating into amino acids and aligning individually in MUSCLE (Edgar, 2004) using default parameters in Geneious. Ribosomal RNA genes were individually aligned in MAFFT (Katoh et al., 2005) using the E–INS–i algorithm and tRNA genes were individually aligned in MUSCLE using default parameters also in Geneious. All these individual alignments were concatenated into a single matrix using SequenceMatrix (Vaidya et al., 2011). We divided the data into a total of 63 data blocks (13 mitochondrial PCGs divided into individual codon positions, 22 tRNAs, and 2 rRNAs) and used PartitionFinder v.2.1.1 (Lanfear et al., 2012) using the "greedy" algorithm (heuristic search) with branch lengths estimated as "linked" to search for the best-fit scheme as well as to estimate the model of nucleotide evolution for each partition. The final matrix consisted of 15,080 aligned bp and 51 taxa.

We performed both maximum likelihood (ML) and Bayesian analyses on the total evidence dataset. For the ML analysis, we used the best-fit partitioning scheme recommended by PartitionFinder with the GTRCAT model applied to each partition and analyzed using RAxML 7.2.8 (Stamatakis et al., 2008). Nodal support was evaluated using 1000 replications of rapid bootstrapping implemented in RAxML. For the Bayesian analysis, we applied a different, unlinked model for each partition, as recommended by PartitionFinder, and ran four runs with four chains each for 70.1 million generations, sampling every 5000 generations in MrBayes 3.2.6 (Ronquist et al., 2012). We plotted the likelihood trace for each run to assess convergence in Tracer 1.6 (Rambaut and Drummond, 2003–2009), and discarded an average of 25% of each run as burn-in. Both analyses were run on XSEDE (Extreme

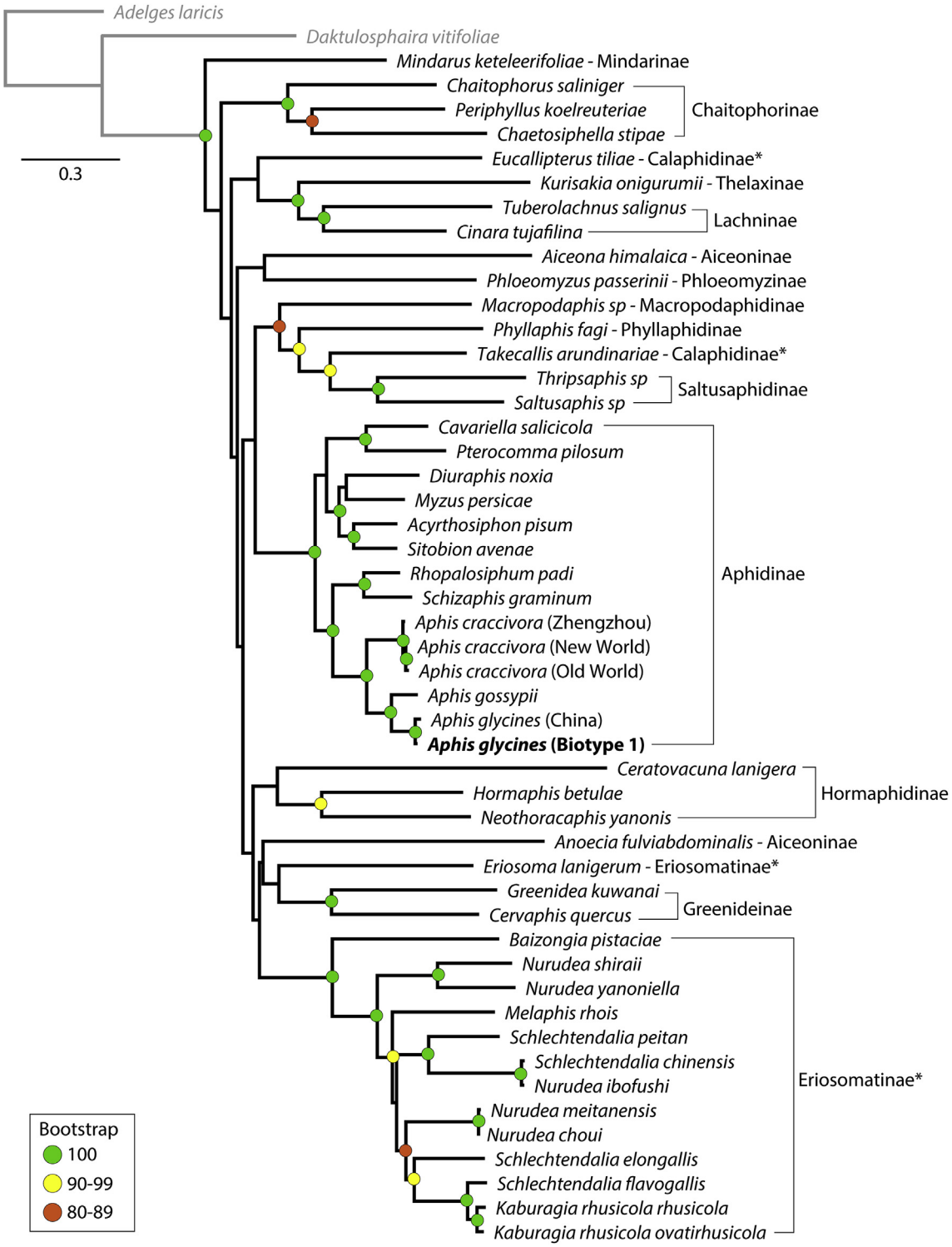

Fig. 3. Mtgenome phylogeny of Aphididae inferred using maximum likelihood. Nodal support values are color coded. The subfamily names with asterisk indicate paraphyletic groups. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Science and Engineering Discovery Environment, https://www.xsede.org) through the CIPRES Science Gateway (Miller et al., 2011). The resulting trees were visualized in Geneious. Our aligned dataset and the resulting trees, as well as all associated data were deposited to Mendeley Data (https://doi.org/10.17632/9s2c2mwhzv.1).

2.4. Comparative analysis of mtgenomes

We first computed the content of A + T for each PCG and ribosomal RNA gene, combined tRNA genes, the control region, and the repeat regions (if present) for each species in Geneious. Then we calculated mean AT percentages for the three gene categories of those coding sequences (ATP synthase subunits, cytochrome oxidase subunits, and NADH dehydrogenase subunits), combined ribosomal RNA sequences,

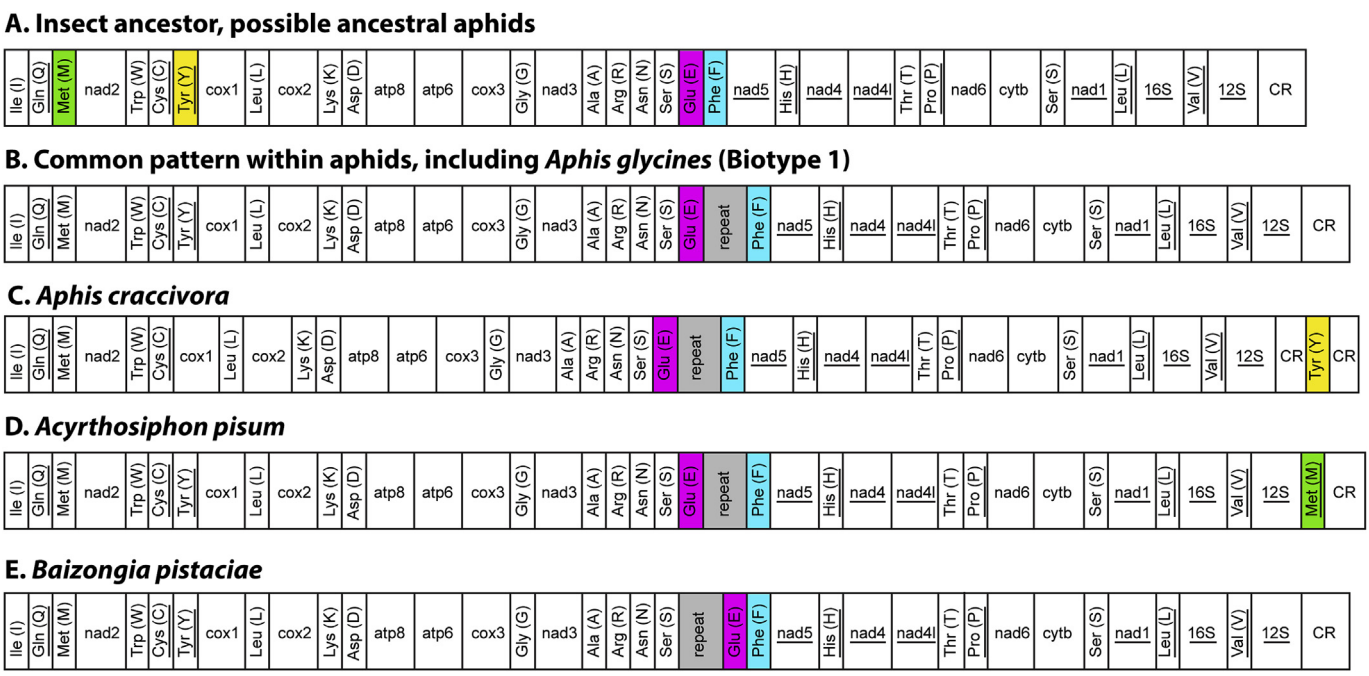

Fig. 4. Five types of mtgenome arrangement known from aphids. tRNA genes that show gene rearrangement patterns from the ancestral genome organization are shown in color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

combined transfer RNA sequences, the control regions, and the repeat regions. We compared the AT percentage in each of these partitions across our taxa to investigate whether there was any lineage specific pattern (Table 1).

To examine possible variations in gene order within Aphididae, we compared the gene arrangement of all available mtgenomes and looked for patterns of gene rearrangement and insertion of repeat regions, which were mapped to the ML tree.

2.5. Phylogeny-based comparative analysis of repeat region

In order to test whether the repeat region is species-specific or is phylogenetically conserved, we reconstructed a phylogeny based on the repeat region sequences. If the repeat region evolved independently in different lineages, the resulting phylogeny using these sequences would have a topology that is incongruent with the mtgenome tree. If the resulting phylogeny recovers relationships similar to the mtgenome tree, it can be inferred that the repeat region evolved prior to the diversification of different aphid lineages and have remained in the mtgenomes and co-diversified along with other mitochondrial genes. We first extracted the sequences of the repeat region from the mtgenome sequences from all species that are known to have the repeat region between tRNA-Glu and tRNA-Phe. We used the Old World representative of *Aphis craccivora* (NC_031387), since there were two mtgenomes available for this species. Although *Baizongia pistaciae* has a repeat region in a different position than the rest of the species (see below, Fig. 5), we included the species because of the similar sequence characteristics. In total, we included the repeat regions of 13 species for this analysis. We aligned the sequences in MAFFT using the E–INS–i algorithm implemented in Geneious, and used RAxML with the GTRCAT model to infer the phylogeny, also in Geneious. Nodal support was evaluated using 1000 replications of rapid bootstrapping.

To test whether the repeat region could have evolved through duplication of the existing control region, we performed pairwise alignments between the repeat region and the control region of the 13 species using MAFFT with the E–INS–i algorithm in Geneious. We recorded the number of identical sites between the repeat region and the control region for each species and calculated pairwise % identity in Geneious. Furthermore, we created an alignment consisting of both the repeat region and the control region of the 13 species, which were used for building a phylogeny to test whether the repeat region would group with the control region of the respective species or not. The alignment was made using MAFFT, and the phylogeny was inferred using RAxML with the GTRCAT model to infer the phylogeny, as described above.

3. Results

3.1. Mitochondrial genome description of the soybean aphid

The complete sequence of the soybean aphid mtgenome was 17,954 bp and this mtgenome contained 13 PCGs, 22 tRNA genes, 2 rRNA genes, as well as a 2322 bp repeat region between tRNA-Glu and tRNA-Phe, and a 1158 bp AT-rich control region between *12S* and tRNA-Ile (Fig. 1 and Table 1). The gene order was identical to the ancestral insect mtgenome (Cameron, 2014b), except for the repeat region, which has been reported from some aphid species (Thao et al., 2004; Wang et al., 2013; Zhang et al., 2016a, 2016b). The overall base composition was highly biased towards adenine and thymine (83.5%) (Table 2). Of the coding regions, the cytochrome oxidase subunits (*COI*, *COII*, *COIII*, *CYTB*) collectively had the lowest AT bias (79.6%), while the NADH dehydrogenase subunits (*ND1-ND6*) had the highest AT bias (86.4%). This mtgenome had several non-coding regions spread throughout (Table 1), and the notable intergenic regions are 57 bp between *ND5* and tRNA-His and 10 bp between tRNA-Ser2 and *ND6*. In the repeat region between tRNA-Glu and tRNA-Phe, a 195 bp unit repeated 11.9 times, representing the highest reported repeats among the known aphid mtgenomes to date. All PCGs started with typical ATN start codons, while three of those genes (*ND2*, *COI*, *ND4*) had a partial stop codon with T, and one gene (*COIII*) had a partial stop codon with TA. The large and small rRNA subunits (*16S* and *12S*) were 1261 bp and 769 bp, respectively, and collectively had an AT bias of 85.4%. All the tRNAs were predicted to have clover-leaf secondary structures except for tRNA-Ser1 of which the dihydrouridine arm formed a 5 bp loop rather than a stem (Fig. 2). All tRNAs had typical anticodons known in other aphids. The control region included a 172 bp unit repeated three times.

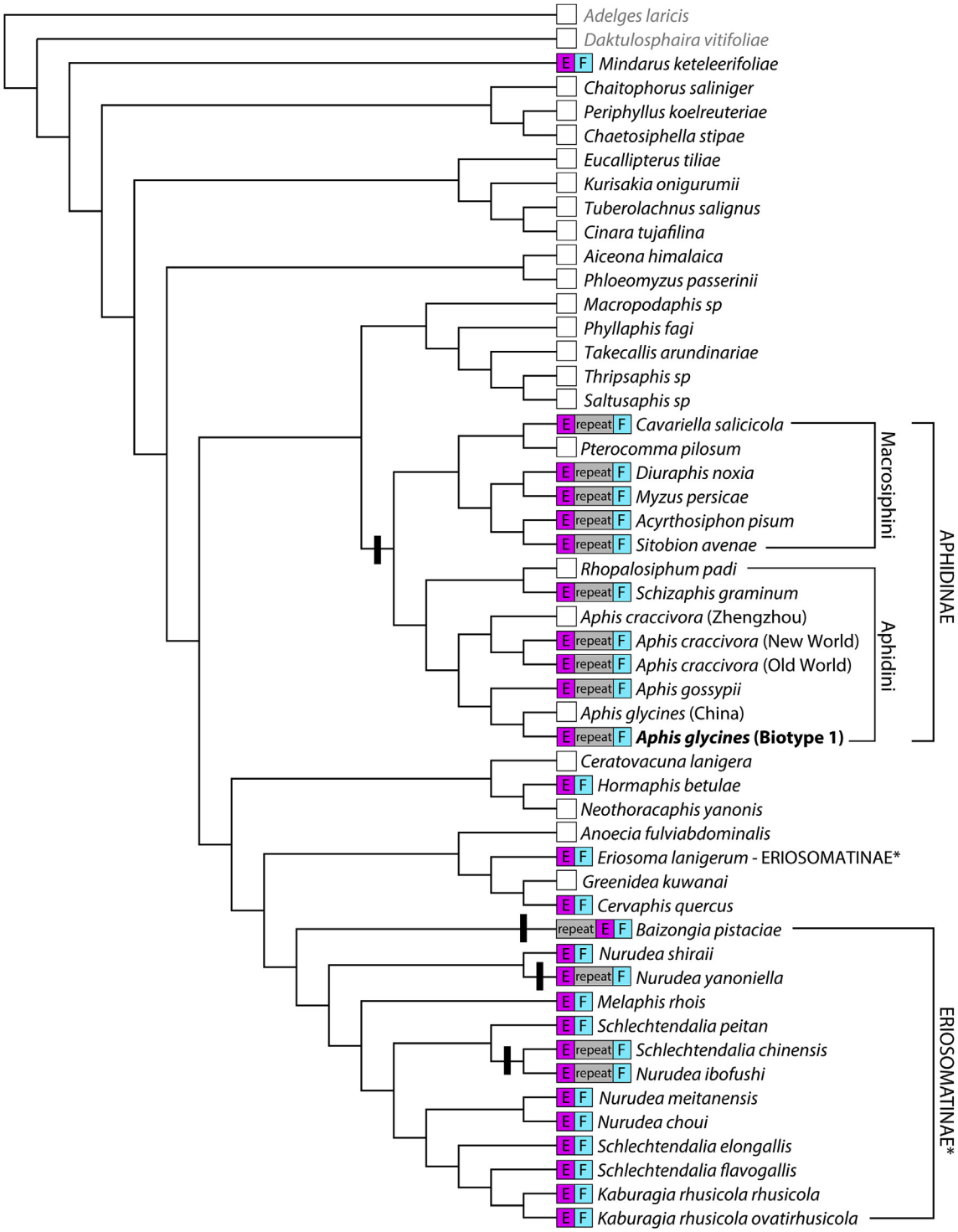

Fig. 5. Evolution of repeat region within Aphididae. The tree shown here is the ML tree converted into a cladogram. tRNA-Glu is shown as a pink box with a letter E, and tRNA-Phe is shown as a blue box with a letter F. When a repeat region is present, it is shown as a gray box. White box indicates unknown data. The black bars on the phylogeny indicate the evolution of repeat region. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.2. Phylogeny of aphididae based on mitochondrial genome sequences

Both ML (Fig. 3) and Bayesian analyses (Fig. S1) recovered Aphididae as monophyletic, and found the subfamily Mindarinae to be the earliest diverging lineage within the family, followed by Chaitophorinae. The internal relationships among the major subfamilies differed between the two analyses, but the nodal support values for the backbone relationships were generally poor (bootstrap values or posterior probability values less than 50) in both analyses. Of the subfamilies that included multiple representatives, Calaphidinae and Eriosomatinae were recovered as paraphyletic, while the other subfamilies were found to be monophyletic, including Lachninae, Saltusaphidinae, Aphidinae, Hormaphidinae, and Greenideinae. In the ML analysis, the Aphidinae was found to be sister to a clade consisting of

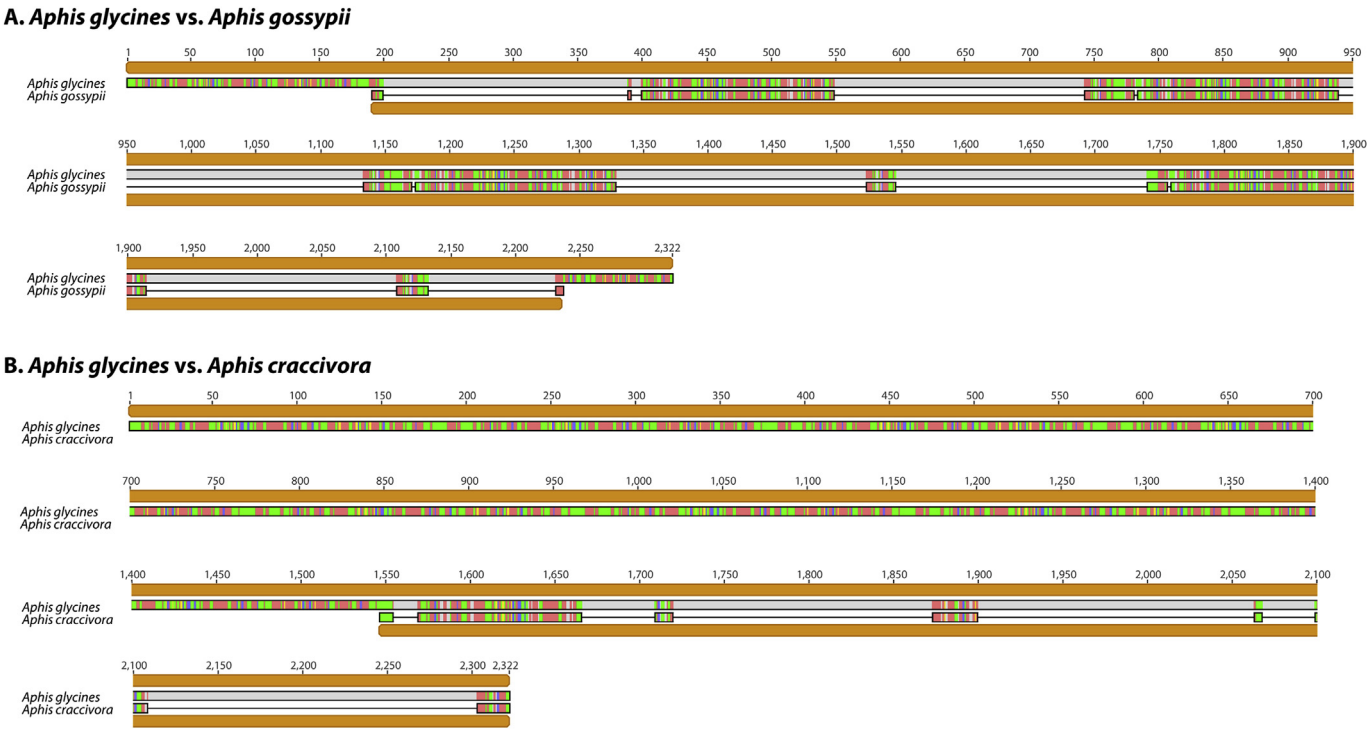

Fig. 6. Pairwise alignment of the repeat region within three *Aphis* species. A. Pairwise alignment between *A. glycines* and *A. gossypii*; B. Pairwise alignment between *A. glycines* and *A. craccivora*.

Macropodaphidinae, Phyllaphidinae, Calaphidinae (*Takecallis arundinariae*) and Saltusaphidinae, but with a low nodal support, and in the Bayesian analysis, it was not possible to determine the affinity of Aphidinae because of the polytomy among the subfamilies. Both analyses recovered an identical topology for the subfamily Aphidinae. Within Aphidinae, the genus *Aphis* was found to be monophyletic, the three species had the following relationship: (*A. craccivora* (*A. gossypii* + *A. glycines*)). The soybean aphid Biotype 1 generated in this study formed a clade with the soybean aphid from China. Overall, the mtgenome data were useful for robustly resolving the relationships within subfamilies, but did not provide sufficient resolutions for determining the relationships among the subfamilies.

3.3. Comparative mitogenomics

Based on the comparisons of available mtgenome sequences, the overall base composition was highly biased towards A and T, ranging between 80.8% and 84.9% (Table 2). The AT percentage of the ATP synthase subunits ranged from 83.1% to 89.6%, with the soybean aphid Biotype 1 having the lowest percentage. The cytochrome oxidase subunits collectively had lower AT bias than other genes, ranging from 76.8% to 81.7%. The NADH dehydrogenase subunits had a narrow range of AT bias, ranging from 85% to 87.8%.

By comparing the 27 complete mtgenome sequences available, we identified at least five types of gene arrangement within Aphididae (Fig. 4). The first type was the ancestral insect mtgenome gene arrangement with 37 genes without any repeat regions or gene rearrangement. This pattern was found in 14 species (Fig. 5), including the most basal lineage within the family, represented by *Mindarus keteleerifoliae* (Mindarinae), as well as *Hormaphis betulae* (Hormaphidinae), *Cervaphis quercus* (Greenideinae), and 10 species of Eriosomatinae. The remaining four types of gene arrangement could be broadly categorized in terms of whether there was an insertion of a repeat region between tRNA-Glu and tRNA-Phe and whether there was a gene rearrangement or duplication (Fig. 5). Of the 13 species with the mtgenomes that deviated from the ancestral mtgenome structure, 12 species had the repeat region in various lengths inserted between tRNA-Glu and tRNA-Phe, which has been proposed as an aphid-specific feature (Wang et al., 2013, 2014). This pattern was found in all known mtgenomes of Aphidinae, including all *Aphis* species, as well as three species of Eriosomatinae (Fig. 5). *Baizongia pistaciae* (Eriosomatinae) also had the repeat region, but it was inserted between tRNA-Ser1 and tRNA-Glu instead of between tRNA-Glu and tRNA-Phe. In terms of gene rearrangement, only the cowpea aphid, *Aphis craccivora*, had tRNA-Tyr rearranged with the strand changed from minor to major, and inserted in the middle of the control region. So, this tRNA was rearranged from its usual place before *COI* to the control region. In terms of gene duplication, the pea aphid (*Acyrthosiphon pisum*) had two copies of tRNA-Met, one in its usual place before *ND2*, and another after *12S* with the strand changed from major to minor.

3.4. Evolution of the repeat region

We mapped the repeat region insertion to the ML phylogeny to test whether they show any phylogenetic patterns (Fig. 5). All members of the subfamily Aphidinae with complete mtgenomes had the repeat region. Within Eriosomatinae, both the ancestral arrangement with no insertion and the inserted repeat region were found. In two occasions, the species with the ancestral arrangement was found to be sister to the species with the repeat region between tRNA-Glu and tRNA-Phe.

When we examined the repeat region between the soybean aphid and *A. gossypii*, we found that the 784 bp repeat region of *A. gossypii* aligned well with that of the soybean aphid, but in nine separate fragments ranging in size from 3 bp to 196 bp (Fig. 6A). The repeat region of *A. craccivora* is much shorter (179 bp). A pairwise alignment of the repeat regions between the soybean aphid and *A. craccivora* showed that the repeat region of *A. craccivora* aligned to the latter half of the repeat region of *A. glycines* and in seven fragments, ranging in size from 5 bp to 97 bp (Fig. 6B).

In addition to mapping the repeat region insertions to the phylogeny, we examined the sequence characteristics of the repeat regions of the 13 species with this insertion. Among the 13 species, the soybean

Table 3
Matrix of shared nucleotide numbers based on the alignment of the repeat region sequences of the 13 species. Each number represents the number of identical sites between each pairwise comparison. The cells are colored coded to create a heatmap, with darker color indicating more shared sites.

	Aphis glycines	*Aphis gossypii*	*Aphis craccivora*	*Schizaphis graminum*	*Acyrthosiphon pisum*	*Sitobion avenae*	*Diuraphis noxia*	*Myzus persicae*	Cavariella salicicola	Baizongia pistaciae	Nurudea yanoniella	Nurudea ibofushi	Schlechtendalia chinensis
Aphis glycines		720	140	382	958	176	403	196	442	222	258	381	377
Aphis gossypii	720		130	298	490	161	327	181	330	193	219	262	260
Aphis craccivora	140	130		77	129	73	95	79	102	90	65	91	91
Schizaphis graminum	382	298	77		396	156	346	177	351	181	225	225	223
Acyrthosiphon pisum	958	490	129	396		199	458	214	489	256	266	391	387
Sitobion avenae	176	161	73	156	199		170	177	170	116	137	127	126
Diuraphis noxia	403	327	95	346	458	170		200	436	205	220	257	257
Myzus persicae	196	181	79	177	214	177	200		187	147	159	153	152
Cavariella salicicola	442	330	102	351	489	170	436	187		208	207	266	264
Baizongia pistaciae	222	193	90	181	256	116	205	147	208		213	221	219
Nurudea yanoniella	258	219	65	225	266	137	220	159	207	213		272	270
Nurudea ibofushi	381	262	91	225	391	127	257	153	266	221	272		666
Schlechtendalia chinensis	377	260	91	223	387	126	257	152	264	219	270	666	

aphid mtgenome had the largest repeat region (2322 bp). The mean length of the repeat region among the 13 species was 690 (± 576) bp, with the smallest repeat region found in *Aphis craccivora* (178–179 bp). In terms of the base composition, this repeat region was highly biased towards AT, ranging between 84.2% and 91.5% (Table 2). The alignment of the repeat regions of the 13 species revealed that this region was highly conserved across the species (Table 3), although there were species-specific length variations. We created a matrix of shared sites, which showed that the repeat region of the soybean aphid had many identical nucleotide sites with the remaining 12 species (Table 3). Notably, 720 sites of the 784 bp repeat region (91.8%) of *A. gossypii* and 140 sites of the 179 bp region (78.2%) of *A. craccivora* were identical to the soybean aphid repeat region. The phylogeny based on the repeat region sequences recovered similar relationships to the mtgenome phylogeny (Fig. 3), although some of the backbone relationships within Aphidinae had low nodal support values (Fig. 7).

The pairwise alignments between the repeat region and the control region revealed that on average 38% of the bases were identical between the two regions (Table 4). The highest percentage of shared sites was found in *Diuraphis noxia* (51.4%), and the lowest was found in *Myzus persicae* (10.8%). Of the 13 species, 3 species (*Aphis glycines, Aphis gossypii, Acyrthosiphon pisum*) had longer repeat regions than the control region, indicating insertions of tandem repeats. The remaining 10 species had shorter repeat regions than the control region, which could have resulted from deletions. The phylogenetic analysis based on the sequences of both regions recovered two clades, one consisting solely of the repeat region sequences and another consisting of the control region sequences (Fig. 8). The internal relationships were similar between the two clades, but the placements of *Sitobion avenae* and *Cavariella salicicola* were different between the two clades.

4. Discussion

This study presents the first sequenced and annotated complete mtgenome of the soybean aphid (Biotype 1). Previously, Wang et al. (2013) generated a partial mtgenome (13,002 bp) of the soybean aphid from China by PCR and primer walking, but it lacked the region spanning tRNA-Ala to *ND5*, because the long AT-rich repeat region prevented the authors from designing appropriate primers to sequence through the complete sequence by primer walking. Our shotgun sequencing approach has enabled us to sequence through this repeat region, and thus allowed us to properly compare the resulting complete mtgenome with other available aphid mtgenomes.

The soybean aphid mtgenome is the largest among all known aphid mtgenomes. Of the 27 available complete mtgenomes, including the new one herein, the size ranges from 15,088 bp (*Hormaphis betulae*) to 17,954 bp (*Aphis glycines* Biotype 1), with the mean size of 15,862 bp. Thus, the size of the soybean aphid mtgenome is unusually large. The only other aphid mtgenome that is larger than 17,000 bp is that of the green peach aphid, *Myzus persicae* (17,382 bp). The large size of the soybean aphid mtgenome is largely due to its large repeat region between tRNA-Glu and tRNA-Phe. Because of this unusual length, we took an additional measure of reassembling this portion of the mtgenome from the raw data and we were able to confirm its validity. In terms of base composition, the soybean aphid mtgenome is 83.5% biased towards A + T, which is within a typical range of what is known in aphids (Wang et al., 2013; Zhang et al., 2013, 2016a, 2016b). The same pattern goes for individual PCGs, rRNAs and tRNAs, as well as the control and repeat region. The A + T content is the highest in the repeat region, followed by the NADH dehydrogenase subunits, tRNAs, rRNAs, control region, and the ATP synthase subunits. The cytochrome oxidase subunits had the lowest A + T content. This pattern is also similar among other aphid mtgenomes. In terms of the mtgenome organization, the soybean aphid mtgenome has all 13 PCGs, 22 tRNAs, 2 rRNAs, and one control region, and only deviates from the ancestral insect mtgenome in that it has the repeat region between tRNA-Glu and

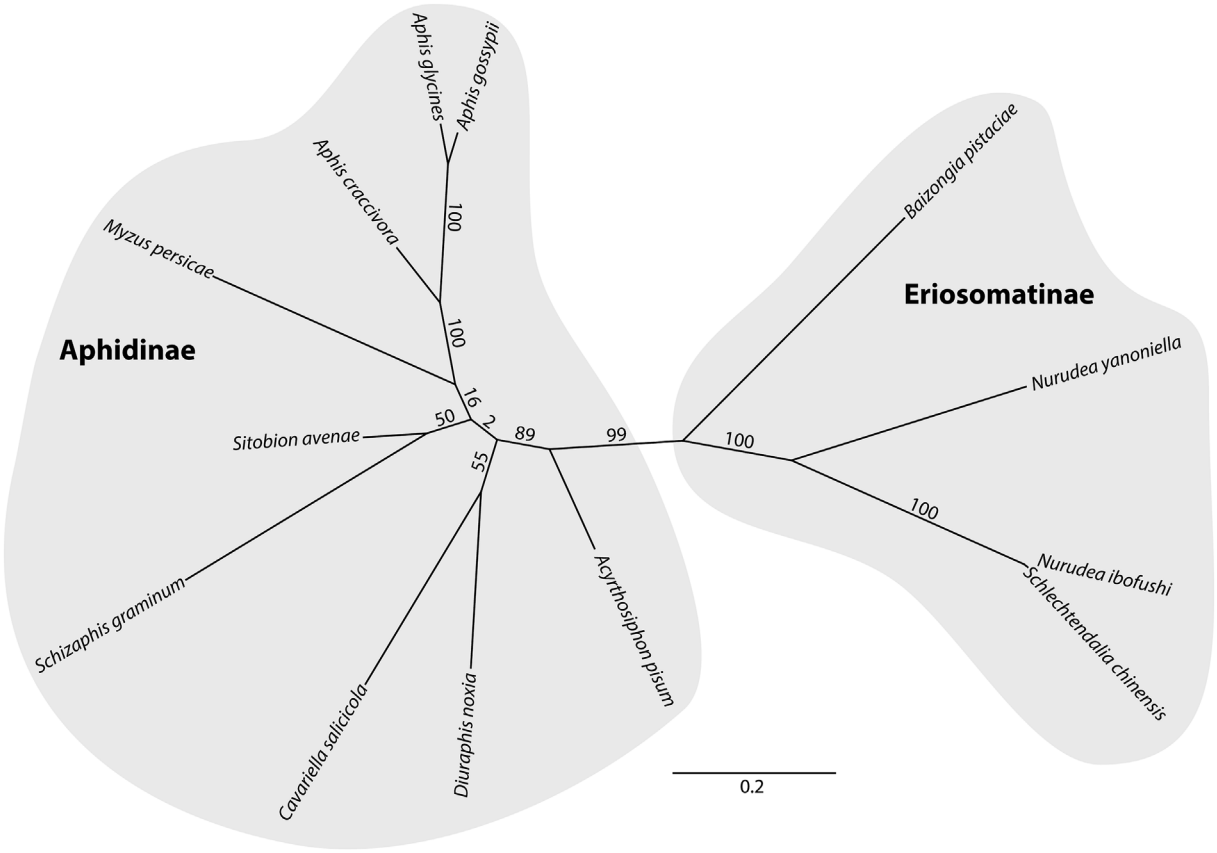

Fig. 7. Unrooted network phylogeny of 13 aphid species with the repeat region between tRNA-Glu and tRNA-Phe. Numbers on the branches are bootstrap support values.

Table 4
Comparison between the control region and the repeat region of the 13 species known to have the repeat region between tRNA-Glu and tRNA-Phe.

	Control Region (bp)	Repeat Region (bp)	# Identical Sites	Pairwise % Identity
Aphis glycines	1158	2322	759	33.5
Aphis gossypii	627	784	332	40.2
Aphis craccivora	645	179	113	17.5
Schizaphis graminum	622	608	308	45.2
Acyrthosiphon pisum	1006	1488	609	40.5
Sitobion avenae	430	261	179	43.2
Diuraphis noxia	664	644	345	51.4
Myzus persicae	2531	307	219	10.8
Cavariella salicicola	1137	702	449	39.5
Baizongia pistaciae	623	395	237	39.7
Nurudea yanoniella	591	452	254	43.4
Nurudea ibofushi	713	670	332	43.4
Schlechtendalia chinensis	705	670	327	42.9
Mean	**881**	**729**	**343**	**37.8**

tRNA-Phe, but this pattern is frequently found within Aphididae (Wang et al., 2013; Zhang et al., 2013, 2016a, 2016b).

We present the most comprehensive mtgenome phylogeny of Aphididae to date, including 49 aphid species. Both ML and Bayesian analyses found the subfamily Mindarinae to be the earliest diverging lineage within the family with strong nodal support. This subfamily is considered a relict group (Chen et al., 2017) and our result is congruent with the previous analyses (Chen et al., 2017; Ren et al., 2017). Also congruent with Chen et al. (2017), we recovered a monophyletic Chaitophorinae to be the second earliest diverging lineage. The subfamily Aphidinae is the most species-rich group within the family, but it has not been clear which group is sister to this subfamily because the two previous studies did not converge on the same relationships. In our phylogeny, we found that Aphidinae was sister to a clade consisting of Macropodaphidinae, Phyllaphidinae, Calaphidinae, and Saltusaphidinae in the ML analysis, similar to the finding of Chen et al. (2017), but in the Bayesian analysis, this topology was not supported. The incongruence between the inference methods suggests that the mtgenome data do not have sufficient phylogenetic information to resolve deep-level relationships among the aphid subfamilies. We think that more conserved markers, such single-copy nuclear protein coding genes, would be necessary to confidently resolve the higher-level relationships within Aphididae. Within Aphidinae, we recovered monophyletic Aphidini, consisting of the genera *Aphis*, *Schizaphis*, and *Rhopalosiphum*, and monophyletic Macrosiphini, consisting of the genera *Cavariella*, *Pterocomma*, *Diuraphis*, *Myzus*, *Acyrthosiphon*, and *Sitobion*. Given that both the ML and the Bayesian analyses recovered the same relationships within Aphidinae, we consider the mtgenome data to be appropriate for resolving phylogenetic relationships at this level.

Our study has analyzed all available aphid mtgenomes, which provides a unique phylogenetic perspective to examine the evolution of the repeat region between tRNA-Glu and tRNA-Phe. Thao et al. (2004) published the first aphid mtgenome (*Schizaphis graminum*), and reported this repeat region, but because the focus of their study was to highlight gene rearrangement in whiteflies, the uniqueness of this repeat region was overlooked. Wang et al. (2013) performed a comparative analysis of five aphid species belonging to Aphidinae, and highlighted this repeat region as special and identified that this region included several tandem repeats. They speculated that this repeat region could be another origin of replication, citing previous studies on other hemipteran mtgenomes (Dotson and Beard, 2001; Li et al., 2011).

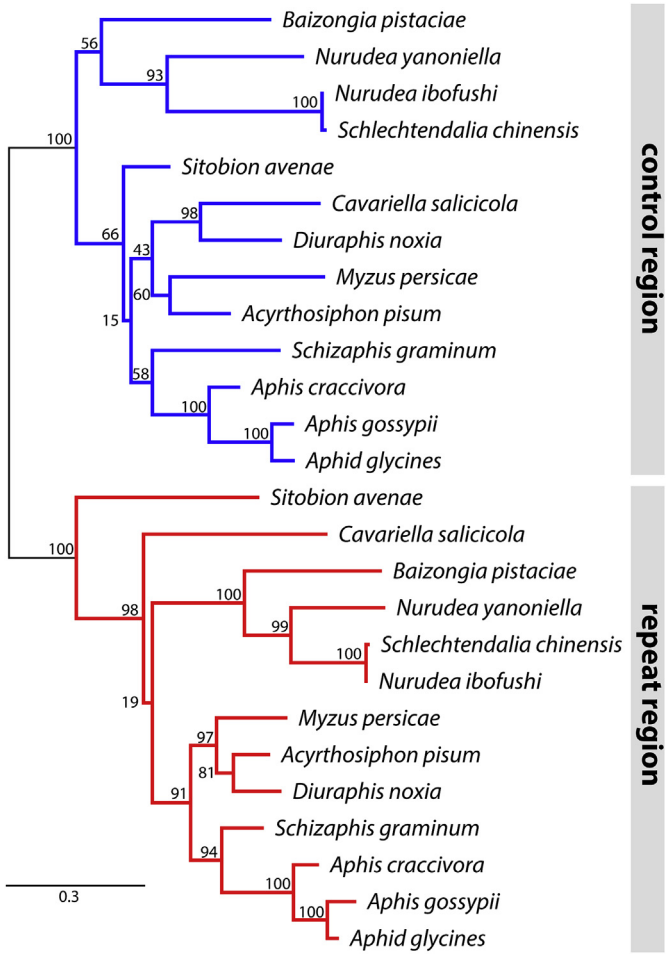

Fig. 8. Phylogeny based on mitochondrial control region and repeat region of 13 aphid species known to have the repeat region between tRNA-Glu and tRNA-Phe. The analysis recovered two clades, one consisting of the control region sequences (blue) and another consisting of the repeat region sequences (red). Numbers on the branches are bootstrap support values. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

As additional mtgenomes of aphid species belonging to the subfamilies other than Aphidinae (Wang et al., 2014, 2015; Zhang et al., 2013) were described, it became clear that some species have the repeat region, while others lack the repeat region. This realization prompted Wang et al. (2014) to hypothesize that this repeat region may be typical of Aphidinae, but not of other aphid subfamilies, which Wang et al. (2015) continued to support with a description of a basal aphid lineage (*Mindarus keteleerifoliae*). Since then, a large number of aphid mtgenomes have been published, but there has not been an updated synthesis of the evolution of the repeat region.

When we map the repeat region on to the phylogeny (Fig. 5), we find that it is a common feature for all species within Aphidinae, but only found in a few species within Eriosomatinae. The most basal aphid species with the complete mtgenome (*Mindarus keteleerifoliae*) does not have this repeat region, which suggests that the ancestral aphids probably had an ancestral mtgenome gene arrangement without any repeat region, and the repeat region originated during the diversification of aphids. If we treat the repeat region as a binary character which can be either present or absent, it is possible to infer that there have been at least four independent evolution events of the repeat region throughout the phylogeny of Aphididae, once in the common ancestor of all Aphidinae, and three independent types within Eriosomatinae (Fig. 5). Kim et al. (2011) estimated the age of Aphidinae to be 67–68 mya, and Ren at al. (2017) estimated it to be 56.6 mya. These divergence time estimates suggest that the repeat region must have persisted in the mtgenomes of Aphidinae for nearly 60 million years. Within Eriosomatinae, *Baizongia pistaciae* is the earliest diverging lineage and Ren et al. (2017) estimated that it diverged from other eriosomatines nearly 90 mya. This species has the repeat region inserted before tRNA-Glu, which is the unique pattern found among the aphids. Interestingly, *Nurudea shiraii* and *N. yanoniella* are sister to each other, but the former does not have the repeat region while the latter does. According to Ren et al. (2017), these two species diverged about 42.5 mya, which suggests that there has been ample time to evolve different mtgenome structures although they share a common ancestor.

Wang et al. (2015) suggested that the repeat region within the family is lineage-specific and occurred from independent evolutionary events. Our method of mapping the repeat region as a binary character onto the phylogeny (Fig. 5) supports this idea, but the actual examination and comparison of the repeat region sequences of the 13 species reveal that the repeat region is more conserved within the aphids and the origin of the repeat region is possibly more ancient than what can be inferred from character mapping. The alignment of the repeat region sequences of the 13 species shows that there are a large number of shared nucleotides among the species, and even between Aphidinae and Eriosomatinae (Table 3). For example, the repeat region of the soybean aphid shares 381 sites of the 670 bp repeat region (56.9%) of *Nurudea ibofushi*. Furthermore, a phylogenetic analysis based on the repeat region sequences shows a topology that is largely congruent with the mtgenome phylogeny (Fig. 7). If the repeat region evolved independently in different species, one would expect either no phylogenetic signal from these sequences or different resulting topologies, but our study shows that there are sites in the repeat region that are phylogenetically conserved. This finding suggests that the common ancestor of Aphidinae and Eriosomatinae, which originated about 142.8 mya (Ren et al., 2017), probably evolved the repeat region once, and there have been numerous events in which the repeat region was lost.

The pairwise comparisons of the repeat region among the three *Aphis* species provide additional insights about the evolution of the repeat region. The finding that the 784 bp repeat region of the cotton aphid aligned well with that of the soybean aphid (Fig. 6A) suggests that the common ancestor between *A. glycines* and *A. gossypii* probably had a repeat region that was about the size of what is currently present in *A. gossypii*. These two species belong to the gossypii group within the genus *Aphis*, and they are separated by about 18 million years of evolution (Kim et al., 2011). However, throughout the diversification of the species within the gossypii group, additional tandem repeats were included to the repeat region of the soybean aphid, making its repeat region longer. It would be interesting to examine the characteristics of the repeat regions among the species in the gossypii group to test whether there is a recognizable pattern of insertions. The cowpea aphid belongs to the craccivora group, which is quite divergent from the gossypii group, separated by about 33 million years (Kim et al., 2011). The shared repeat region between the soybean aphid and *A. craccivora* is much shorter (Fig. 6B) and this is because the repeat region of *A. craccivora* is only 179 bp in length, which could be due to deletions in the repeat region.

Looking at the broad pattern across the aphids, our analysis points to a hypothesis that the repeat region is actually an ancient feature within aphids that has persisted for nearly 150 million years, and that there have been a number of losses and insertions throughout the diversification, giving rise to the current pattern. Because of the energy generating function of mitochondria, it is often hypothesized that mtgenomes of animals are under selection for small size and are without non-coding regions other than the control region (Rand, 1993; Zhang and Hewitt, 1997; Sayadi et al., 2017). Therefore, the persistence of the repeat region in aphids, although it has been lost in several species, begs the question about its origin and function.

Most insect species maintain the ancestral mitochondrial gene arrangement with a single AT-rich control region with many tandem repeats between 12S rRNA and the tRNA-Ile (Boore, 1999; Cameron, 2014b). This region is often called the control region or the origin of replication because it is thought to be responsible for initiation of mtDNA transcription and replication (Clayton, 1991; Saito et al., 2005). The tandem repeats often found in this non-coding region are considered to be a result of slipped-strand mispairing during mtDNA replication (Levinson and Gutman, 1987; Zhang and Hewitt, 1997). The repeat region between tRNA-Glu and tRNA-Phe found in aphids also contain tandem repeats, which prompted some researchers to hypothesize that it could be the second origin of replication (Wang et al., 2015), but this idea has not been empirically tested.

Among vertebrates, some species of fishes, amphibians, reptiles and birds are known to have tandem duplications comprising the control region with adjacent genes, thus have two control regions (reviewed in Urantowka et al., 2018). It is hypothesized that having two control regions can provide selective advantage because they allow more efficient initiation of replication or transcription (Jiang et al., 2007). At least in birds, duplication of the mitochondrial control region is associated with increased longevity (Skujina et al., 2016). Therefore, it is conceivable to hypothesize that duplication of the control region could be a mechanism that gave rise to the repeat region in aphids. To test this idea, we performed pairwise alignments between the control region and the repeat region of the 13 aphid species. If there is a high sequence similarity between the two, then it is reasonable to assume that the repeat region is basically a duplicated control region. Indeed we did find various levels of sequence similarities between the two regions in all 13 species (Table 4), which points to a possibility that the repeat region could have been a duplicated control region. However, based on the fact that only about 38% of the nucleotide sequences are identical between the two regions, we can speculate that the duplication event must have been a very ancient event and the differences between the two regions can be attributed to subsequent accumulation of mutations and indels. The phylogenetic analysis of both regions of the 13 species recovered two clades, each corresponding to the control region and the repeat region (Fig. 7). This sort of pattern can only be achieved if the gene duplication event predated the divergence of the 13 species. In other words, the repeat region must have originated at least in the common ancestor of all 13 species and followed its own evolutionary trajectory independent of the control region.

The presence of a non-coding repeat region other than the control region has been reported in a few insect species (reviewed in Sayadi et al., 2017), and this region has also been referred to as a tandem repeat unit (TRU) (Wan et al., 2012) or a long intergenic spacer (LIGS) (Sayadi et al., 2017). Because most insects do not have such a repeat region, the presence of the repeat region is often reported as a curious exception along with various speculations about its function. Recently, Sayadi et al. (2017) reported two long repeat regions located between *NAD1* and tRNA-Ser, and *NAD2* and tRNA-W in 4 species of seed beetles in the genus *Callosobruchus* and *Acanthoscelides*. These repeat regions were highly variable in size (114–10,408 bp), making these insects with the largest known mtgenomes. Interestingly, they reported that pairwise comparisons of these repeat regions among the four seed beetle species did not yield sequence similarities, which contrasts with our finding that even the distantly related aphids share some sequence similarities (Table 3). To explain the function of these long repeat regions, Sayadi et al. (2017) offered an evolutionary hypothesis based on the metabolic biology of these insects. The mitochondrial respiration of the adult seed beetles mostly relies on the fatty acids stored during the larval period, which reduces the need for building complex I of the oxidative phosphorylation pathway (NADH dehydrogenase). They suggested that the repeat regions might serve to economize the translational machinery of the mtgenome by reducing polycistronic transcription and/or posttranscriptional modification.

Many aphid species show alternation of generation and are parthenogenetic (Moran, 1992; Simon et al., 2002), and a single female can produce an enormous number of offspring. This process must be energetically demanding, which can benefit from efficient mitochondrial machinery powered by two origins of replication. Although it is unknown whether the repeat region indeed functions as the second origin of replication, the ancient origin of the repeat region, as inferred in this study, might be related to the antiquity of cyclical parthenogenesis (Moran, 1992). Certainly, many species have completely lost the repeat region and are still capable of cyclical parthenogenesis, which means that the repeat region is not a necessity for this unique mode of reproduction. Alternatively, we can look into the feeding biology of aphids to see if the efficient mitochondrial machinery can be advantageous. Many pest aphids rely on feeding on plants with complex chemistries that likely include oxidizing molecules (Guerrieri and Digilio, 2008) these in turn could negatively affect mitochondrial function (Yan et al., 1997). The up regulation of mitochondrial replication and/or apoptosis (recycling) would likely be a critical resource for aphids to cope with oxidative damage, and if the repeat region indeed serves as a second origin of replication, the maintenance of this feature in the mtgenome can confer adaptive significance to the aphids. Indeed, more work is required to understand the function of this repeat region.

In conclusion, the soybean aphid mtgenome represents the largest known mtgenome within Aphididae, and it has the phylogenetically conserved repeat region between tRNA-Glu and tRNA-Phe. In all other aspects, this species has a similar mtgenome compared to other mtgenomes within Aphidinae. The genus *Aphis* is the most species-rich genus within Aphididae (Kim et al., 2011), as more *Aphis* mtgenomes become available analyses will be needed to determine if the patterns observed in this study will remain consistent.

Acknowledgments

Funding for the sequencing of the genome of *Aphis glycines* Biotype 1 was provided by the United Soybean Board (USB), Illinois Soybean Association (ISA), and the North Central Soybean Research Program (NCSRP) to R. Giordano.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmb.2019.103208.

References

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69, 313–319.

Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27, 1767–1780.

Cameron, S.L., 2014a. How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. Syst. Entomol. 39, 400–411.

Cameron, S.L., 2014b. Insect mitochondrial genomics: implications for evolution and phylogeny. Annu. Rev. Entomol. 59, 95–117.

Cameron, S.L., Barker, S.C., Whiting, M.F., 2006. Mitochondrial genomics and the new insect order Mantophasmatodea. Mol. Phylogenet. Evol. 38, 274–279.

Cameron, S.L., Lambkin, C.L., Barker, S.C., Whiting, M.F., 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.

Cameron, S.L., Miller, K.B., D'Haese, C.A., Whiting, M.F., Barker, S.C., 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Athropoda). Cladistics 20, 534–557.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinf. 13, 238.

Chen, J., Wang, Y., Jiang, L., Qiao, G., 2017. Mitochondrial genome sequences effectively reveal deep branching events in aphids (Insecta: Hemiptera: Aphididae). Zool. Scripta 46, 706–717.

Clayton, D.A., 1991. Replication and transcription of vertebrate mitochondrial DNA. Annu. Rev. Cell Biol. 7, 453–478.

Dixon, A.F.G., 1997. Aphid Ecology an Optimization Approach. Springer Science & Business Media.

Dotson, E.M., Beard, C.B., 2001. Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*. Insect Mol. Biol. 10, 205–215.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Favret, C., 2018. Aphid Species File. Version 5.0/5.0. [10/28/2018]. http://Aphid.SpeciesFile.org.
Fenn, J.D., Song, H., Cameron, S.L., Whiting, M.F., 2008. A mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. Mol. Phylogenet. Evol. 49, 59–68.
Guerrieri, E., Digilio, M.C., 2008. Aphid-plant interactions: a review. J. lant Interact. 3, 223–232.
Jiang, Z.J., Castoe, T.A., Austin, C.C., Burbrink, F.T., Herron, M.D., McGuire, J.A., Parkinson, C.L., Pollock, D.D., 2007. Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. BMC Evol. Biol. 7, 123.
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.
Kim, C.S., Schaible, G., Garret, G., Lubowski, R., Lee, D., 2008. Economic impacts of the U.S. soybean aphid infestation: a multi-regiional competitive dynamic analysis. Agric. Resour. Econ. Rev. 37, 227–242.
Kim, H., Lee, S., Jang, Y., 2011. Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. PLoS One 6, e24749.
Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.
Li, H., Gao, J., Liu, H., Liu, H., Liang, A., Zhou, X., Cai, W., 2011. The architecture and complete sequence of mitochondrial genome of an assassin bug *Agriosphodrus dohrni* (Hemiptera: reduviidae). Int. J. Biol. Sci. 7, 792–804.
Li, Y.Q., Chen, J., Qiao, G.X., 2017. Complete mitochondrial genome of the aphid *Hormaphis betulae* (mordvilko) (Hemiptera: Aphididae: Hormaphidinae). Mitochondrial DNA Part A 28, 265–266.
Miller, M.A., Holder, M.T., Vos, R., Midford, P.R., Liebowitz, T., Chan, L., Hoover, P., Warnow, T., 2011. The CIPRES Portals. CIPRES. http://www.phylo.org/sub_sections/portal.
Moran, N.A., 1992. The evolution of aphid life cycles. Annu. Rev. Entomol. 37, 321–348.
Ragsdale, D.W., Landis, D.A., Brodeur, J., Heimpel, G.E., Desneux, N., 2011. Ecology and management of the soybean aphid in North America. Annu. Rev. Entomol. 56, 375–399.
Ragsdale, D.W., Voegtlin, D.J., O'neil, R.J., 2004. Soybean aphid biology in North America. Ann. Entomol. Soc. Am. 97, 204–208.
Rambaut, A., Drummond, A.J., 2003-2009. Tracer: MCMC Trace Analysis Tool Version v1.5.0.
Rand, D.M., 1993. Endotherms, ectotherms, and mitochondrial genome size variation. J. Mol. Evol. 37, 281–295.
Ren, Z., Harris, A.J., Dikow, R.B., Ma, E., Zhong, Y., Wen, J., 2017. Another look at the phylogenetic relationships and intercontinental biogeography of eastern Asian – North American *Rhus* gall aphids (Hemiptera: Aphididae: Eriosomatinae): evidence from mitogenome sequences via genome skimming. Mol. Phylogenet. Evol. 117, 102–110.
Ren, Z.M., Bai, X., Harris, A.J., Wen, J., 2016. Complete mitochondrial genome of the *Rhus* gall aphid *Schlechtendalia chinensis* (Hemiptera: Aphididae: Eriosomatinae). Mitochondrial DNA Part B: Resources 1, 849–850.
Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.
Saito, S., Tamura, K., Aotsuka, T., 2005. Replication origin of mitochondrial DNA in insects. Genetics 171, 1695–1705.
Sayadi, A., Immonen, E., Tellgren-Roth, C., Arnqvist, G., 2017. The evolution of dark matter in the mitogenome of seed beetles. Genome Biol. Evol. 9, 2697–2706.
Simon, J.-C., Rispe, C., Sunnucks, P., 2002. Ecology and evolution of sex in aphids. Trends Ecol. Evol. 17, 34–39.
Skujina, I., McMahon, R., Lenis, V.P.E., Gkoutos, G.V., Hegarty, M., 2016. Duplication of the mitochondrial control region is associated with increased longevity in birds. Aging 8, 1781–1789.
Song, N., An, S., Yin, X., Cai, W., Li, H., 2016a. Application of RNA-seq for mitogenome reconstruction, and reconsideration of long-branch artifacts in Hemiptera phylogeny. Sci. Rep. 6, 33465.
Song, N., Zhang, H., Li, H., Cai, W., 2016b. All 37 mitochondrial genes of aphid *Aphis craccivora* obtained from transcriptome sequencing: implications for the evolution of aphids. PLoS One 11, e0157857.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 75, 758–771.
Sun, W., Huynh, B.L., Ojo, J.A., Coates, B.S., Kusi, F., Roberts, P.A., Pittendrigh, B.R., 2017. Comparison of complete mitochondrial DNA sequences between old and new world strains of the cowpea aphid, *Aphis craccivora* (Hemiptera: Aphididae). Agri Gene 4, 23–29.
Taanman, J.W., 1999. The mitochondrial genome: structure, transcription, translation and replication. Biochim. Biophys. Acta Bioenerg. 1410, 103–123.
Thao, M.L., Baumann, L., Baumann, P., 2004. Organization of the mitochondrial genomes of whiteflies, aphids, and psyllids (Hemiptera, Sternorrhyncha). BMC Evol. Biol. 4, 25.
Urantowka, A.D., Kroczak, A., Silva, T., Padron, R.Z., Gallardo, N.F., Blanch, J., Blanch, B., Mackiewicz, P., 2018. New insight into parrots' mitogenomes indicates that their ancestor contained a duplicated region. Mol. Biol. Evol. 35, 2989–3009.
Vaidya, G., Lohman, D.J., Meier, R., 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27, 171–180.
Wang, Y., Chen, J., Jiang, L.Y., Qiao, G.X., 2015. The complete mitochondrial genome of *Mindarus keteleerifoliae* (Insecta: Hemiptera: Aphididae) and comparison with other Aphididae insects. Int. J. Mol. Sci. 16, 30091–30102.
Wang, Y., Huang, X.L., Qiao, G.X., 2013. Comparative analysis of mitochondrial genomes of five aphid species (Hemiptera: Aphididae) and phylogenetic implications. PLoS One 8, e77511.
Wang, Y., Huang, X.L., Qiao, G.X., 2014. The complete mitochondrial genome of *Cervaphis quercus* (Insecta: Hemiptera: Aphididae: Greenideinae). Insect Sci. 21, 278–290.
Wolstenhome, D.R., 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. 141, 173–216.
Yan, L.-J., Levine, R.L., Sohal, R.S., 1997. Oxidative damage during aging targets mitochondrial aconitase. Proc. Natl. Acad. Sci. 94, 11168–11172.
Zhang, B., Ma, C., Edwards, O., Fuller, S., Kang, L., 2013. The mitochondrial genome of the Russian wheat aphid *Diuraphis noxia*: large repetitive sequences between trnE and trnF in aphids. Gene 533, 253–260.
Zhang, B., Zheng, J., Liang, L., Fuller, S., Ma, C.S., 2016a. The complete mitochondrial genome of *Sitobion avenae* (Hemiptera: Aphididae). Mitochondrial DNA 27, 945–946.
Zhang, D.-X., Hewitt, G.M., 1997. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. Biochem. Syst. Ecol. 25, 99–120.
Zhang, S., Luo, J., Wang, C., Li, C., Jiang, W., Cui, J., Rajput, L.B., 2016b. Complete mitochondrial genome of *Aphis gossypii* glover (Hemiptera: Aphididae). Mitochondrial DNA Part A 27, 854–855.
Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics 29, 2669–2677.