Deep-level phylogeny of Cicadomorpha inferred from mitochondrial genomes sequenced by NGS

Nan Song1, Wanzhi Cai2 & Hu Li2

Recent development and advancement of next-generation sequencing (NGS) technologies have enabled the determination of mitochondrial genome (mitogenome) at extremely efficiency. In this study, complete or partial mitogenomes for 19 cicadomorphan species and six fulgoroid species were reconstructed by using the method of high-throughput sequencing from pooled DNA samples. Annotation analyses showed that the mitogenomes obtained have the typical insect mitogenomic content and structure. Combined with the existing hemipteran mitogenomes, a series of datasets with all 37 mitochondrial genes (up to 14,381 nt total) under different coding schemes were compiled to test previous hypotheses of deep-level phylogeny of Cicadomorpha. Thirty-seven species representing Cicadomorpha constituted the ingroup. A taxon sampling with nine species from Fulgoroidea and six from Heteroptera comprised the outgroup. The phylogenetic reconstructions congruently recovered the monophyly of each superfamily within Cicadomorpha. Furthermore, the hypothesis (Membracoidea + (Cicadoidea + Cercopoidea)) was strongly supported under the heterogeneous CAT model.

The Cicadomorpha are generally regarded as an infraorder of Hemiptera. This insect group includes three superfamilies: Membracoidea (leafhoppers and treehoppers), Cicadoidea (cicadas) and Cercopoidea (spittlebugs and froghoppers), with approximately 30,000 described species. The monophyly of the entire Cicadomorpha and of each superfamily has never been questioned based on morphological characters and/or molecular evidence. Internal relationships within Cicadomorpha have been addressed by several studies, which attempted to assess higher hemipteran or auchenorrhynchan phylogeny and thus included representatives of superfamilies of Cicadomorpha. In particular, the higher-level relationships within Cicadomorpha were definitively resolved by the multi-locus, quantitative analyses of Cryan (2005) and Cryan & Urban (2012). Currently, most Hemiptera/Auchenorrhyncha systematists already agree that the arrangement of (Membracoidea + (Cicadoidea + Cercopoidea)) is the major relationship within Cicadomorpha.

Leston et al. (1954) proposed the infraorder Cimicomorpha. After that, Evans (1963) recovered a sister-group relationship between Cicadoidea and Membracoidea based mainly on the head morphological characters. Hamilton (1981) placed Cicadoidea as an ancient lineage within Cicadomorpha and as a sister group of (Cercopoidea + Membracoidea), also based on the evidence from head morphology. In the subsequent molecular analyses based on 18S rDNA data, the hypothesis of (Cicadoidea + (Cercopoidea + Membracoidea)) was further supported. However, employing the similar 18S rDNA sequences, Campbell et al. (1995) recovered Cicadoidea and Cercopoidea as a clade, collectively sister to Membracoidea. The hypothesis of (Membracoidea + (Cicadoidea + Cercopoidea)) was supported by several morphology-based researches. Fossil evidence also supported the grouping of Cicadoidea with Cercopoidea. In the study by Cryan (2005), the same inter-superfamily relationship within Cicadomorpha as those in Campbell et al. (1995) was supported by the combination analyses of three nuclear markers (i.e. 18S rDNA, 28S rDNA and histone 3). Furthermore, the topological arrangement of (Membracoidea + (Cicadoidea + Cercopoidea)) was reinforced by expanding data from seven gene regions of nuclear and mitochondrial markers. Therefore, the phylogenetic hypothesis of (Membracoidea + (Cicadoidea + Cercopoidea)) is the major relationship within Cicadomorpha.

1College of Plant Protection, Henan Agricultural University, Zhengzhou, 450002, China. 2Department of Entomology, China Agricultural University, Beijing, 100094, China. Correspondence and requests for materials should be addressed to N.S. (email: songnan@henau.edu.cn)
coidea + (Cicadoidea + Cercopoidea)) has been overwhelmingly supported by recent contributions as described above. This study aimed to take a phylogenomic approach based on additional mitochondrial genome (mitogenome) data to test relationships among superfamilies within Cicadomorpha.

Mitogenome is one of the most extensively utilized marker in phylogenetic studies of insects. This class of organelle genome usually has high copy numbers, for example, each human cell contains between $10^3$ and $10^4$ copies of the mitogenome. This characteristic makes mtDNA easy to be determined. Complete mitogenome contains 37 mitochondrial genes, which are known to harbor higher substitution rates. They are considered to be well-suited for resolving phylogenies at different taxonomic levels. At present, compared with the whole genome data, mitogenome sequencing allows for much larger scale sampling. Especially in recent years, the advent of next-generation sequencing (NGS) techniques have revolutionized the ease with which mitogenome data can be obtained time- and also cost-effectively for large taxon sampling of insects.

In the following study, we determined 25 hemipeteran insect mitogenome sequences including 19 cicadomorphan species and six fulgoroid species by using an approach of next-generation sequencing from mixed DNA samples. In addition, we combined with the published hemipeteran mitogenome sequences to reconstruct the deep-level phylogeny of Cicadomorpha, with the goal of investigating relationships among superfamilies and families of this group.

Materials and Methods

Ethics Statement. No specific permits were required for the insect specimens collected for this study in China. These specimens were collected in the suburban fields of Xinyang, China. The field studies did not involve endangered or protected species. All sequenced insects are common species in China, and are not included in the “List of Protected Animals in China”.

Taxon sampling. A total of 25 taxa were selected for mitogenome sequencing, with emphasis on the superfamily relationships within Cicadomorpha and the possible sister group lineage of Cicadomorpha (i.e. Fulgoroidea). Specimen identification were conducted by checking adult morphological characters, and by blasting in online identification tool of BOLD systems (Barcode Of Life Database: http://www.boldsystems.org - Identification section) and by the standard nucleotide BLAST in NCBI. The primary specimen materials can be accessed by the Entomological Museum of Henan Agricultural University. The detailed classification information and voucher numbers of species sequenced in this study are listed in Table S1.

In addition, we included all the current available mitogenomes of Membracoidea, Cicadoidea, Cercopoidea and Fulgoroidea in the GenBank (up to September 2016). In total, the ingroup included 37 taxa from Cicadomorpha: 17 species representing Membracoidea, eight species representing Cicadoidea, and 12 species representing Cercopoidea.

Outgroup choice included a taxon sampling based on the results of recent molecular studies. Nine species representing the superfamily of Fulgoroidea served as the close outgroups. Six species representing Heteroptera were selected as the relatively distant outgroups. The complete list of taxa included in this study is given in Table S1.

DNA extraction. Total genomic DNA was isolated from each 95–100% ethanol preserved specimen individually using the TIANamp Micro DNA Kit (TIANGEN BIOTECH CO., LTD) following the manufacturer’s protocol. DNA concentration was measured by Nucleic acid protein analyzer (QUAWELL TECHNOLOGY INC.).

Mitogenome reconstruction. The assembly strategy of complete mitogenome is largely identical to that of Gillett et al. (2014). The minor differences lied in the universal primers designed to amplify bait sequences as those in Song et al. (2016). Additionally, the Illumina HiSeq X Ten sequencing platform was utilized in the present study. Similar amounts of genomic DNA for each individual were pooled to improve the sequencing efficiency for every species. The amount of pooled DNA was quantified at 1.5 μg. Moreover, species with phylogenetically distant relations were mixed in a library to avoid the chimera formation. Totally, three libraries were constructed using Illumina TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Genomic DNA were fragmented with a Covaris sonicator to an average insert size of 350 bp. The subsequent de novo genome sequencing was conducted on a single lane of Illumina HiSeq X Ten by Beijing Novogene Bioinformatics Technology Co., Ltd (China). Approximately 10 Gb paired-end reads of 150 bp length were generated for each library. FastQC was used for quality control of raw sequence data to remove reads containing adapters and ploy-N, or low quality reads from the raw data. All the downstream analyses were based on clean data of high quality (avg. Q20 > 90%, and avg. Q30 > 85%). No less than 8 Gb high-quality reads for each library were used in de novo assembly using IDBA-UD v. 1.1.1, with default settings.

The information of bait sequences blasted against each mitochondrial contig determined are presented in Table S2. Only the assembled contigs to which at least one Sanger sequence could be matched with certainty were retained for further analysis. Mapping to identified mitochondrial contigs were performed using BWA v. 0.7.13 under default parameters. Mapping statistics were obtained with Qualimap and Tablet, in order to check the quality of the assembled contigs.

The preliminary annotation for the contigs identified by bait sequences were conducted using MITOS, with default settings and the invertebrate genetic code for mitochondria. The resultant gene boundaries were further checked and corrected by alignment against 20 published mitogenome sequences from Cicadomorpha and Fulgoroidea (see details in Table S1). Furthermore, hand alignment checking was also conducted by blasting each predicted gene for every mitogenome against GenBank data in Web BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), to ensure the identity of the mitochondrial genes identified.
Sequence alignment. The nucleotide sequences of each protein-coding gene were aligned based on codons using the invertebrate mitochondrial genetic code in the Perl script TransAlign. Each of tRNA and rRNA genes was aligned using MAFFFT (version 7) under iterative refinement method incorporating the most accurate local (E-INS-I) pairwise alignment information. Alignments were checked in MEGA 6. Gaps were stripped by Gap Strip/Squeeze v2.1.0 with 40% Gap tolerance (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html). Finally, all alignments were concatenated to construct two matrices using FASconCAT_v1.0.5, one including RNA genes (i.e. PCG), and one excluding RNA genes (i.e. PCGRNA).

A part from two datasets mentioned above, a series of datasets were compiled by the following methods to test the influences of recoding schemes on the phylogenetic estimate: (1) PCG AA: translating nucleotides of protein-coding genes into amino acid sequences; (2) PCGDegen: 13 protein-coding genes were re-coded by degenerating all sites including synonymous substitutions to IUPAC ambiguity codes through Degen v1.4 (Degen-code)46,49; (3) PCGDegenRNA: 13 protein-coding genes with Degen-coding combined with 24 RNA genes. Additionally, the datasets of PCG and PCGRNA were masked by using Alicut version 2.050,51, combined with Alicut version 2.050,51. Finally, to investigate the potential long-branch effect on the tree topology, a series of reduced datasets (i.e. 43 taxa datasets without the outgroup Fulgoroidea) were created to further phylogenetic analyses.

Sequence saturation in the combined protein-coding genes and RNA genes were assessed using the index of substitution saturation (Iss)42 as implemented in the DAMBE 553, respectively. Estimates of nonsynonymous (dN) and synonymous (dS) substitution rates of concatenated protein-coding genes were obtained by the method of Yang and Nielsen (2000)52 using the program yn00 as implemented in PAML 4.9. The One-way Analysis of Variance (ANOVA) is performed in Excel 2016.

Phylogenetic analyses. Phylogenetic reconstructions were based on the mitogenome sequences of the full datasets with 52 taxa and the reduced datasets with 43 taxa under maximum likelihood (ML) and Bayesian inference (BI). Prior to ML analyses, PartitionFinder56 was employed to infer the optimal partitioning strategy. Simultaneously, the Bayesian Information Criterion (BIC) were used to choose the best models for the combined nucleotide and amino acid datasets under a greedy search with RAxML57,58. The data blocks were defined by gene types (each of 13 PCGs, 22 tRNAs and two rRNAs) and by codon positions (each of three codon positions for PCGs), totally 63 independent blocks were employed for the datasets of 52taxa_PCGRNA, 52taxa_PCGDegenRNA, 43taxa_PCGRNA, and 43taxa_PCGDegenRNA. The partition schemes and best-fit models selected for each dataset are provided in Table S3. Gene types and codon positions cannot be distinguished when data were masked, thus no partition analyses were applied to these datasets (i.e. Alicut-52taxa_PCG, Alicut-52taxa_PCGRNA, Alicut-43taxa_PCG, and Alicut-43taxa_PCGRNA).

ML analyses were conducted using IQ-TREE9 as implemented on the multicore version of IQ-TREE 1.5.5. The partition schemes and best-fit models selected by PartitionFinder were applied to corresponding dataset. Branch support was estimated using Ultrafast option for bootstrap analysis, with 1000 replicates. The detailed commands are as following: iqtree-omp -s dataset.nex -st DNA (or AA) -bb 1000 -alrt 1000.

Bayesian analyses were performed using a parallel version of PhyloBayes (pb_mpi1.5a)59,60 as implemented on a HP server with twenty-four CPU and 320 Gi memory. For all PhyloBayes analyses, two independent runs were performed, and started from random topology, respectively. Each run implemented two differentially heated chains, with at least 30,000 cycles. The CAT-GTR model was used for nucleotide analyses, while the CAT model for amino acids. Convergence was monitored using tracecomp (“maxdiff” value < 0.1) and tracecomp (minimum effective size > 100). A 25% burn-in was applied after checking for stationarity, and a consensus tree was calculated.

For each phylogenetic analysis, we utilized FigTree v1.4.365 to visualize the consensus tree and the corresponding branch lengths. The one-way ANOVA analyses for branch lengths of major groups are performed in Excel 2016. The bootstrap supports (BS) of ≥75 and posterior probabilities (PP) of ≥0.95 were considered to be credible support values for tree nodes. All sequence alignment files and tree files built in this article are available in the Treebase: http://purl.org/phylo/treebase/phylows/study/TB2:S19876.

Hypothesis testing. To test the statistical significance of alternative hypotheses of Cicadomorpha, we compared all possible relationships among three superfamilies, namely (Membracoidea + (Cicadoidea + Cercopoida)), (Cicadoidea + (Membracoidea + Cercopoida)), and (Cercopoida + (Membracoidea + Cicadoidea)). The topology tests were conducted using the datasets with the full 52 taxa (i.e. 52taxa_PCG, 52taxa_PCGDegen, 52taxa_PCG_AA, 52taxa_PCGRNA, 52taxa_PCGDegenRNA, Alicut_52taxa_PCG, and Alicut_52taxa_PCGRNA). The site-log-likelihood values were calculated under the GTR + I + G model for nucleotides and the MtREV + I + G model for amino acids using TREE-PUZZLE 5.340 (the command-line option: -wsl). The obtained values were used as input for the software CONSEL84. Constraint likelihood trees were constructed based on the basis of dataset of 52taxa_PCGRNA with RAxML57,58 using the model GTRGAMMA and using partitions selected by PartitionFinder. Three competing hypotheses were statistically tested among each other by AU75, KH86, SH87, WKH and WSH.

Results Assembly of mitogenomes. Twenty-five hemipteran insect mitogenomes determined were identified from the individual contigs by bait sequences. No chimeric formation was found when we inspected the base coverage along each mitogenome with the software Tablet32. Because every site in each mitochondrial contig corresponded to an identical nucleotide. The statistics from the program BWA38 showed that read coverage for all gene regions in each contig was no less than 30-fold. Mean coverage for every mitogenome varied between 103-fold
and 1,922-fold. The average depth of coverage was 475-fold, with 17 contigs ranging from 220-fold to 742-fold coverage and four contigs at >1,000-fold coverage (Fig. 1). Although with the higher sequence coverage, complete mitogenomes (including the full 37 mitochondrial genes and the entire control region) were identified for seven insect species. The remainder were nearly complete (12 mitogenomes including the full 37 mitochondrial genes and the partial control region) or partial mitogenomes (six mitogeneonomes including 20–35 mitochondrial genes and no partial control region). The lengths of the mitogenome sequences ranged from 8,455 nt to 16,226 nt, of which 17 mitogenomes had sequences’ length more than 15,033 nt, four ones ranged from 12,231 nt to 14,803 nt, and the rest had sequences <10,000 nt. For the incomplete mitogenomes, the missing segments were mainly located adjacent to the putative control region. The complete or nearly complete mitogenomes have the consistent gene content and organization with other published auchenorrhynchan insect mitogenomes. All new mitogenome sequences have been deposited in GenBank (accession numbers are presented in Table S1).

The results of the substitution saturation tests based on the concatenated protein-coding genes with 52 taxa showed that the values of substitution saturation index (Iss) for the first and/or second codons and all sites of PCG were significantly smaller than the critical values (Iss.cSym or Iss.cAsym). However, the Iss values of the third codon positions were larger than the Iss.cSym and Iss.cAsym (Table 1). This result indicated that the third codon positions might have a negative impact on phylogenetic analysis. Removal of the long-branched outgroups (i.e. Fulgoromorpha) did not significantly reduced the saturation degree. There still showed saturation in the third codon positions and in the RNA genes only (Table 1). From the point of view of the lower Iss values, data masking reduced the degree of saturation.

To explore the correlation between sequence evolution rate and tree topology, we compared the nonsynonymous (dN) and synonymous (dS) evolutionary rates for each sequence pair within five groups base on the concatenated protein-coding genes (Table 2). The one-way ANOVA analysis revealed incongruence in the dN values between groups ($P = 0.0429$). This result was mainly due to the lower dN values of Cercopoidea. When we omitted the Cercopoidea to rerun one-way ANOVA analysis, there was no significant difference across all

---

**Table 1.** Saturation test performed by DAMBE. Note: The occurrences of saturation are indicated in bold.

| Gene regions     | NumOTU | Iss   | Iss.cSym | P   | Iss.cAsym | P   |
|------------------|--------|-------|----------|-----|-----------|-----|
| 52taxa_PCG1      | 32     | 0.529 | 0.808    | 0.000 | 0.553     | 0.0303 |
| 52taxa_PCG2      | 32     | 0.398 | 0.808    | 0.000 | 0.553     | 0.0000 |
| 52taxa_PCG1 + 2  | 32     | 0.458 | 0.814    | 0.000 | 0.570     | 0.0000 |
| 52taxa_PCG3      | 32     | 0.821 | 0.808    | 0.1088 | 0.553     | 0.0000 |
| 52taxa_PCG123    | 32     | 0.558 | 0.818    | 0.000 | 0.572     | 0.0000 |
| 52taxa_Alicut-PCG| 32     | 0.449 | 0.815    | 0.000 | 0.571     | 0.0000 |
| 52taxa_RNA       | 32     | 0.705 | 0.807    | 0.000 | 0.550     | 0.0000 |
| 43taxa_PCG1      | 32     | 0.487 | 0.808    | 0.000 | 0.553     | 0.0000 |
| 43taxa_PCG2      | 32     | 0.351 | 0.808    | 0.000 | 0.553     | 0.0000 |
| 43taxa_PCG1 + 2  | 32     | 0.416 | 0.814    | 0.000 | 0.570     | 0.0000 |
| 43taxa_PCG3      | 32     | 0.806 | 0.808    | 0.8142 | 0.553     | 0.0000 |
| 43taxa_PCG123    | 32     | 0.524 | 0.818    | 0.000 | 0.572     | 0.0000 |
| 43taxa_Alicut-PCG| 32     | 0.394 | 0.815    | 0.000 | 0.571     | 0.0000 |
| 43taxa_RNA       | 32     | 0.658 | 0.807    | 0.000 | 0.550     | 0.0000 |

---

**Figure 1.** The graph of mean sequencing coverage versus contig length for 25 identified mitogenomic assemblies.
| Group            | Higher taxa       | Species               | $dN$  | $dS$  | $dN/dS$ |
|------------------|-------------------|-----------------------|-------|-------|---------|
| 1                | Cercopoidea       | Abidama producta      | 0.1875| 4.5129| 0.0416  |
| 1                | Cercopoidea       | Aeneolamia contigua   | 0.1869| 4.5579| 0.0410  |
| 1                | Cercopoidea       | Aphrophora abti       | 0.1942| 4.4320| 0.0438  |
| 1                | Cercopoidea       | Aphrophora sp.        | 0.1893| 4.6337| 0.0409  |
| 1                | Cercopoidea       | Callitetix biformis   | 0.1900| 4.6523| 0.0408  |
| 1                | Cercopoidea       | Callitetix brasconoides| 0.1897| 4.5338| 0.0419  |
| 1                | Cercopoidea       | Callitetix sp.        | 0.1900| 4.5334| 0.0419  |
| 1                | Cercopoidea       | Callitetix versicolor | 0.1910| 4.5573| 0.0419  |
| 1                | Cercopoidea       | Cosmoscarta bispecularis| 0.2131| 4.7267| 0.0451  |
| 1                | Cercopoidea       | Diceroprocta semicincta| 0.1943| 4.6352| 0.0419  |
| 2                | Cicadoidea        | Gaeana maculata       | 0.2100| 4.6076| 0.0456  |
| 2                | Cicadoidea        | Magicicula tredecim   | 0.2998| 4.6179| 0.0649  |
| 2                | Cicadoidea        | Meimuna opalifera     | 0.2605| 4.5730| 0.0570  |
| 2                | Cicadoidea        | Platyleura kaempferi  | 0.2114| 4.7246| 0.0447  |
| 2                | Cicadoidea        | Pomponia linearis     | 0.2039| 4.6080| 0.0442  |
| 2                | Cicadoidea        | Tettigades auripilosa | 0.2475| 4.6274| 0.0535  |
| 2                | Cicadoidea        | Tettigades ulnaria    | 0.2247| 4.7728| 0.0471  |
| 2                | Cicadoidea        | Diceroctena seminuncata| 0.1943| 4.6352| 0.0419  |
| 3                | Fulgoroidea       | Lycorma delicatula    | 0.2849| 4.7153| 0.0604  |
| 3                | Fulgoroidea       | Lydda sp.              | 0.2465| 4.6355| 0.0532  |
| 3                | Fulgoroidea       | Nilaparvata sp.       | 0.2265| 4.8451| 0.0467  |
| 3                | Fulgoroidea       | Oliarias sp.           | 0.2122| 4.6527| 0.0456  |
| 3                | Fulgoroidea       | Pentastiridius sp.     | 0.2056| 4.6321| 0.0444  |
| 3                | Fulgoroidea       | Scolops sp.            | 0.2186| 4.6690| 0.0468  |
| 3                | Fulgoroidea       | Staloka damnosus       | 0.2278| 4.6292| 0.0492  |
| 3                | Fulgoroidea       | Staloka sp.            | 0.2048| 4.6000| 0.0445  |
| 3                | Fulgoroidea       | Sogatella fuscifera    | 0.2134| 4.7127| 0.0453  |
| 4                | Heteroptera       | Diplonychus rusticus   | 0.2184| 4.6572| 0.0452  |
| 4                | Heteroptera       | Enithares tribus       | 0.2160| 4.5846| 0.0471  |
| 4                | Heteroptera       | Hypanopis longirostris| 0.2154| 4.5687| 0.0472  |
| 4                | Heteroptera       | Lygus hesperus         | 0.2475| 4.6338| 0.0534  |
| 4                | Heteroptera       | Macrocytus gibbus      | 0.2578| 4.6158| 0.0558  |
| 4                | Heteroptera       | Yemmalisus parallelus  | 0.2134| 4.7320| 0.0451  |
| 5                | Membracoidea      | Alobaldia tobae       | 0.1929| 4.4817| 0.0430  |
| 5                | Membracoidea      | Darthula hardwickii   | 0.1927| 4.5987| 0.0419  |
| 5                | Membracoidea      | Drabescoides nuchalis  | 0.2143| 4.5832| 0.0468  |
| 5                | Membracoidea      | Empoasca vitia        | 0.2206| 4.6509| 0.0474  |
| 5                | Membracoidea      | Exitana indicus       | 0.2110| 4.6075| 0.0458  |
| 5                | Membracoidea      | Homalodisca vitripennis| 0.2207| 4.6315| 0.0477  |
| 5                | Membracoidea      | Illitina sp.           | 0.2587| 4.6322| 0.0559  |
| 5                | Membracoidea      | Leptobius gazella      | 0.2065| 4.5969| 0.0623  |
| 5                | Membracoidea      | Nephotettix cincticeps| 0.2855| 4.7262| 0.0604  |
| 5                | Membracoidea      | Norvelina sp.          | 0.2093| 4.6171| 0.0453  |
| 5                | Membracoidea      | Oligiana sp.           | 0.2122| 4.6698| 0.0454  |
| 5                | Membracoidea      | Orosius orientalis     | 0.2226| 4.6732| 0.0476  |
| 5                | Membracoidea      | Philotettix sp.        | 0.2291| 4.6404| 0.0494  |
| 5                | Membracoidea      | Tricentrus sp.         | 0.2061| 4.6091| 0.0447  |
| 5                | Membracoidea      | Tricentrus sp.1        | 0.2344| 4.6922| 0.0500  |
| 5                | Membracoidea      | Typhlocyba sp.         | 0.2279| 4.6335| 0.0492  |
| 5                | Membracoidea      | Yamaocephalus yanonis  | 0.2184| 4.6406| 0.0471  |

Table 2. Estimation of nonsynonymous (\(dN\)) and synonymous (\(dS\)) substitution rates by \(yn00\) implemented in PAML.
remaining lineages. There was no significant difference between major groups in the $dS$ values ($P = 0.0539$) or in the $dN/dS$ values ($P = 0.0697$).

**Phylogenetic analyses.** ML analyses based on the full taxa datasets (i.e. 52taxa_PCG, 52taxa_PCGDegen, 52taxa_PCGRNA, 52taxa_PCG_AA, Alicut_52taxa_PCG, and Alicut_52taxa_PCGRNA) consistently resulted in a paraphyletic Cicadomorpha, with respect to the nested position of Fulgoroidea (Fig. 2). Monophyly of each superfamily within Cicadomorpha was very strongly supported (BP = 100). The sister group relationship between Cicadoidea and Cercopoidea was favored with strong nodal support (BP ≥ 85) in all analyses. Table 3 provides the nodal supports and branch lengths for major lineages in each tree. Comparing the support values of deep nodes in each tree, we found that data treatment decreased the support for the inter-superfamily relationships. Specially, the dataset PCGDegenRNA showed the lowest support for the affinity of Cicadoidea with Cercopoidea. This suggested synonymous sites may provide information for the deep-level relations within Cicadomorpha. The one-way ANOVA analysis showed no significant incongruence in the branch lengths between major groups ($P > 0.05$) in the ML analyses. At the family level, the Membracidae were congruently supported as a monophyletic lineage (BP = 100). However, the Cicadellidae was recovered as a paraphyletic assemblage, owing to the close affinity of Olidiana sp. to the clade (Aetalionidae + Membracidae).

Within Cercopoidea, both families Aphrophoridae and Cercopidae were supported as the non-monophyletic groups.

Bayesian analyses based on the full taxa datasets under the heterogeneous model provided the distinct tree topology from those in the ML analyses. All Bayesian analyses recovered the monophyly of Cicadomorpha (Fig. 3), except for the analysis based on the 52taxa_PCG_AA. The Cicadoidea formed a sister group to the Cercopoidea, with posterior probability 1 in the analyses of 52taxa_PCGRNA, 52taxa_PCGDegenRNA, Alicut_52taxa_PCG, and Alicut_52taxa_PCGRNA. The Membracoidea was sister to the clade (Cicadoidea + Cercopoidea). The datasets with recoding schemes displayed the weaker power in resolving the deep-level phylogeny of Cicadomorpha. The branching pattern of (Membracoidea + (Cicadoidea + Cercopoidea)) was not retrieved in the Bayesian trees.
from PCGDegen and PCG_AA. The dataset PCGDegen showed the lowest nodal support for the sister group relation of Cicadoidea + Cercopoidea. The one-way ANOVA analysis revealed significant incongruence in the branch lengths between major groups ($P < 0.05$) in the Bayesian analyses. When removing the Fulgoroidea and Membracoidea to rerun the ANOVA analysis, there was no difference between Cicadoidea and Cercopoidea. This result demonstrated that the potential long-branch effect was introduced by the Fulgoroidea and Membracoidea.

### Taxon deletion experiment.

The taxon deletion experiments were conducted to test for the effect of long-branch attraction. For that, a series of ML analyses were rerun based on the 43 taxa datasets. As a result, all ML analyses with 43 taxa datasets yielded strong evidence for a monophyletic Cicadomorpha ($BP = 100$), with an inter-superfamily relationship of (Membracoidea + (Cicadoidae + Cercopoidea)). The deep nodes defining superfamily relationships received strong bootstrap support values ($BP \geq 78$).

Similarly, the majority of Bayesian analyses based on the 43 taxa datasets reproduced the superfamily relationships within Cicadomorpha as ML analyses based on the 43 taxa datasets. With the exception of 43taxa_PCGDegenRNA, all 43 taxa datasets showed Cicadoidea as sister group to Cercopoidea, and together they were sister to Membracoidea. In the Bayesian tree of 43taxa_PCGDegenRNA, Membracoidea was placed as sister group to Cercopoidea, while Cicadoidea was recovered as an independent lineage within Cicadomorpha.

### Hypothesis testing.

Various tests consistently supported the arrangement of (Membracoidea + Cicadoidae + Cercopoidea) as the best tree topology, regardless of data coding schemes applied (Table 4). In contrast, all tests but for those on the ALICUT_52taxa_PCGDegenRNA significantly rejected the hypothesis of (Cicadoidea + Membracoidea + Cercopoidea)). Additionally, the comparison of topologies clearly rejected the alternative hypothesis of (Cercopoidea + (Membracoidea + Cicadoidae)).

### Discussion

**Efficiency of reconstructing complete mitogenomes.** The present study demonstrates the utility of next-generation sequencing of mixed DNA samples for reconstructing hemipteran mitogenomes. This method
shows a great improvement in the efficiency of achieving a large number of mitogenome data, compared with the traditional Sanger sequencing via primer walking. We used this method to successfully reconstruct 19 cicadomorphian insect mitogenomes and six fulgoroid insect mitogenomes. The number of mitogenome with full-length sequence is relatively small, with only seven ones. For the remaining nearly complete or partial mitogenomes, the missing segments are mainly located in the control region and/or including the adjacent regions. The low capture specificity of the control region could be attributed to the following reasons: 1) the complex nucleotide motifs found in this region (e.g. A+T-rich elements); 2) non-uniform read coverage along the mitogenome determined, especially with the significant drops occurred in the segments corresponding to the control region. Both factors may lead to the sequencing failure and lower assembly efficiency of this specific region. According to previous studies, the current sequencing depth should be sufficient to cover the whole mitogenome. To some extent, more taxon sampling can be added to the pools to improve the sequencing efficiency. Therefore, in order to ensure the completeness of mitogenome, how to increase the sequencing depth for the particular regions of genome and how to assemble the complete mitogenome by developing more efficient assembler will be other important issues.

Outgroup selection for phylogenetic analysis of Cicadomorpha. Outgroup choice is critical in deep phylogenetic studies, because it determines the polarity of characters analyzed. Problematic tree rooting has caused conflict results in the previous studies on the phylogeny of Cicadomorpha. Cryan (2005) discussed the issues on the selection of outgroup in the phylogenetic reconstruction of Cicadomorpha. Despite uncertainties remain on the relationships between Cicadomorpha and their allies, Cryan (2005) thought that it was appropriate to utilize the Fulgoroidea as the outgroup in his study. Besides, he suggested that analysis of Cicadomorpha

Figure 3. Bayesian tree inferred from the dataset of 52 taxa PCGRNA using PhyloBayes under the site-heterogeneous CAT-GTR model. Node numbers show posterior probability values (≥0.95). Scale bar represents substitutions/site. Asterisks designate the species newly sequenced in this study. Other Bayesian trees reconstructed in this paper can be available from the TreeBase: http://purl.orgphylo/treebase/phylows/study/TB2:S19876.
phylogeny should include representatives from Heteroptera as outgroups\(^{10}\). In the current study, we applied an outgroup taxon sample including Heteroptera and Fulgoroidea to the tree reconstruction of Cicadomorpha. Unfortunately, long branch lengths shared by outgroup Fulgoroidea and ingroup Membracoidea were found in the current mitogenome data (Table 3). Moreover, evolutionary rate analyses indicated the distinct values of non-synonymous substitution across five major groups analyzed. These results made us suspect that incongruence between analyses of two inference methods (i.e. ML analysis under homogeneous model and Bayesian analysis under heterogeneous model) from the full datasets might be the consequence of long-branch effect. Excluding long-branch outgroup Fulgoroidea resulted in a congruent result from both inference methods. Cicadomorpha was recovered as a monophyletic group, with the internal relationship of (Membracoidea + (Cicadoidea + Cercopoidea)). Therefore, the Heteroptera may be a more suitable outgroup choice to the Cicadomorpha phylogeny estimate on the current mitogenome data, due to the similar evolutionary rate shared by them.

### Deep-level phylogeny within Cicadomorpha.

For the superfamily relationship of (Membracoidea + (Cicadoidea + Cercopoidea)) within Cicadomorpha, Boulard (1991a,b)\(^{13,14}\) listed two character states supporting this hypothesis: (1) structures of the alimentary canal; (2) the larval behavior of applying excreted liquid. Liang & Fletcher (2002)\(^{15}\) proposed the common antennal features shared by Cicadoidea and Cercopoidea. Rakitov (2002)\(^{16}\) listed structural characters of brachosomes and proteinaceous particles secreted by glandular regions of the Malpighian tubules as synapomorphic for the grouping (Cicadoidea + Cercopoidea). In addition, fossil researches proved the close affinity of Cicadoidea with Cercopoidea\(^{17}\). Recent molecular studies\(^{4,11}\), based mainly on the nuclear gene fragments, also recovered a solid branch pattern of (Membracoidea + (Cicadoidea + Cercopoidea)). Results presented in the current study are the first molecular investigation to utilize mitogenome data only for reconstructing deep-level phylogeny of Cicadomorpha. The majority of our analyses recovered the topology of (Membracoidea + (Cicadoidea + Cercopoidea)), corroborating earlier studies with new data. Nevertheless, the causes for another recovery of (Cicadoidea + (Membracoidea + Cercopoidea)) by only one analysis based on 43taxa_PCGDegenRNA under PhyloBayes may be complex. It is possible that a combined effect of Degen-coding and additional RNA genes contributed to the latter tree structure.

Taken together, the deep-level phylogeny of Cicadomorpha was explored using mitogenome data, with various coding strategies and different algorithms. Our results validate the power of mitogenome for resolving the relationships at the superfamily level in Cicadomorpha. Although the number of mitogenomes available for Cicadomorpha has been almost doubled by this study, phylogenetic result is still preliminary for this megadiverse insect group. In particular, analysis of family level relationship within Cicadomorpha will require more taxon sampling. The related research project is being carried out by the authors.

### References

1. Dietrich, C. H. Evolution of Cicadomorpha (Insecta, Hemiptera). *Denisia* 4, 155–170 (2002).
2. Hamilton, K. G. A. Morphology and evolution of the rynchotan head (Insecta: Hemiptera, Homoptera). *Can. Ent.* 113, 953–974 (1981).
3. Emeuánov, A. F. The phylogeny of the Cicadina (Homoptera, Cicadina) based on comparative morphological data. *Trudy Vsesoy. Entomol. Obshch.* 69, 19–109 (1987).

### Table 4. Tree topology testing. Note: Bold indicates the values of \(p < 0.05\). Mem: Membracoidea, Cic: Cicadoidea, Cer: Cercopoidea.

| Datasets      | Hypothesis      | \(-\ln L\) likelihood | AU | KH | SH | WKH | WSH |
|---------------|-----------------|------------------------|----|----|----|-----|-----|
| 52taxa_PCG    | Mem + (Cic + Cer) | −327098.5              | 1.00 | 1.00 | 1.00 | 1.00 | 1.00|
|               | Cic + (Mem + Cer) | −327209.6              | 0.00 | 0.00 | 0.002 | 0.000 | 0.001|
|               | Cer + (Mem + Cic) | −327199.5              | 0.00 | 0.00 | 0.001 | 0.000 | 0.000|
| 52taxa_PCGDegen| Mem + (Cic + Cer) | −135979.4              | 0.997 | 0.997 | 1.000 | 0.992 | 1.000|
|               | Cic + (Mem + Cer) | −136046.9              | 0.007 | 0.008 | 0.019 | 0.008 | 0.012|
|               | Cer + (Mem + Cic) | −136039.7              | 0.004 | 0.003 | 0.028 | 0.003 | 0.006|
| 52taxa_PCGDegen| Mem + (Cic + Cer) | −170376.6              | 0.998 | 0.998 | 1.000 | 0.998 | 1.000|
|               | Cic + (Mem + Cer) | −170459.9              | 0.002 | 0.002 | 0.007 | 0.002 | 0.003|
|               | Cer + (Mem + Cic) | −170477.7              | 0.000 | 0.000 | 0.002 | 0.000 | 0.000|
| 52taxa_PCGRNA | Mem + (Cic + Cer) | −412235.2              | 1.000 | 0.998 | 1.000 | 0.998 | 1.000|
|               | Cic + (Mem + Cer) | −412338.4              | 0.000 | 0.002 | 0.006 | 0.002 | 0.002|
|               | Cer + (Mem + Cic) | −412353.9              | 0.000 | 0.000 | 0.001 | 0.000 | 0.000|
| 52taxa_PCGDegenRNA| Mem + (Cic + Cer) | −219796.2              | 0.994 | 0.989 | 1.000 | 0.989 | 1.000|
|                | Cic + (Mem + Cer) | −219863.1              | 0.009 | 0.011 | 0.038 | 0.011 | 0.019|
|                | Cer + (Mem + Cic) | −219883.2              | 0.001 | 0.001 | 0.006 | 0.001 | 0.001|
| ALICUT_52taxa_PCG| Mem + (Cic + Cer) | −232954.9              | 0.980 | 0.966 | 0.999 | 0.966 | 0.998|
|                | Cic + (Mem + Cer) | −233040.1              | 0.038 | 0.034 | 0.084 | 0.034 | 0.058|
|                | Cer + (Mem + Cic) | −233064.7              | 0.003 | 0.004 | 0.026 | 0.004 | 0.008|
| ALICUT_52taxa_PCGRNA| Mem + (Cic + Cer) | −281229.9              | 0.897 | 0.890 | 0.982 | 0.890 | 0.979|
|                | Cic + (Mem + Cer) | −281292.7              | 0.112 | 0.110 | 0.206 | 0.110 | 0.171|
|                | Cer + (Mem + Cic) | −281387.9              | 0.001 | 0.000 | 0.006 | 0.000 | 0.001|
4. Sorensen, J. T., Campbell, B. C., Gill, R. J. & Steffen-Campbell, J. D. Non-monophyly of Auchenorrhyncha (Homoptera), based upon 18S rDNA phylogeny: eco-evolutionary and cladistic implications within pre-Heteropterodea Hemiptera (s.l.) and a proposal for new monophyletic suborders. *Pan-Pacific Ent.* 71, 31–60 (1995).

5. von Dohlen, C. D. & Moran, N. A. Molecular phylogeny of the Homoptera: a paraphyletic taxon. *J Mol. Evol.* 41, 211–23 (1995).

6. Campbell, B. C., Steffen-Campbell, J. D., Sorensen, J. J. & Gill, R. J. Paraphyly of Homoptera and Auchenorrhyncha inferred from 18S rDNA nucleotide sequences. *Syst. Entomol.* 20, 175–194 (1995).

7. Scherbakov, D. E. Origin and evolution of the Auchenorrhyncha as shown in the fossil record. Studies on Hemipteran Phylogeny (ed. by Schaefer, C. W.) *Entomol. Soc. Am. Lanham* 31–45 (1996).

8. Yoshizawa, K. & Saigusa, T. Phylogeographic analysis of paraneopteran orders (Insecta: neoptera) based on forewing base structure, with comments on monophyly of auchenorrhyncha (hemiptera). *Syst. Entomol.* 26, 1–13 (2001).

9. Bourquin, T. & Campbell, B. C. Inferring a phylogeny for Hemiptera: falling into the autapomorphic trap. *Denisia* 4, 67–82 (2002).

10. Crampton-Platt, A., Song, N., Liang, A. & Bu, C. A molecular phylogeny of Hemiptera: Cicadidae, Cercopidae and Membracoidea: adding evidence to the controversy. *Syst. Entomol.* 30, 563–574 (2005).

11. Crampton-Platt, A. J. A high-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? *Syst. Entomol.* 37, 7–21 (2012).

12. Lepton, D., Pendergast, J. G. & Southwood, T. R. E. Classification of the Terrestrial Heteroptera (Geocorisae). *Nature* 174, 91–92 (1954).

13. Evans, J. W. The phylogeny of the Homoptera. *Ann. Rev. Entomol.* 8, 77–94 (1963).

14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 20, 173–184 (1995).

15. García-Alcalde, F. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 20, 173–184 (1995).

16. Hamilton, K. G. A. Cretaceous Hemiptera from Brazil: implications for classification. Studies on Hemipteran Phylogeny (ed. by Schaefer, C. W.) *Entomol. Soc. Am. Lanham* 46–64 (1996).

17. Blocker, H. D. Origin and radiation of the Auchenorrhyncha. Studies on Hemipteran phylogeny (ed. by Schaefer, C. W.) *Entomol. Soc. Am. Lanham* 89–110 (1996).

18. Cameron, S. L. Insect mitochondrial genomes, implications for evolution and phylogeny. *Ann. Rev. Entomol.* 59, 95–117 (2014).

19. Rooney, J. P. et al. PCR based determination of mitochondrial DNA copy number in multiple species. *Methods Mol Biol.* 1241, 23–38 (2015).

20. Cui, J. P. & Kocher, T. D. Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Ecol. Evol.* 19, 394–398 (1999).

21. Ma, C., Liu, C., Yang, P. & Käng, L. The complete mitochondrial genomes of two band-winged grasshoppers, Gastrimargus marmoratus and Oedaleus asiaticus. *BMC Genomics* 10, 156 (2009).

22. Sheffield, N. C., Song, H., Cameron, L. & Whiting, M. F. A Comparative Analysis of mitochondrial genomes in Coleoptera (Arthropoda: Insecta) and genome descriptions of six new beetles. *Mol. Biol. Evol.* 25, 2499–2509 (2008).

23. Sheffield, N. C., Song, H. J., Cameron, S. L. & Whiting, M. F. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. *Syst. Biol.* 58, 381–394 (2009).

24. Pons, J., Ribera, I., Bertranpetit, J. & Balke, M. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Mol. Phylogenet. Evol.* 56, 796–807 (2010).

25. Timmermans, M. J. et al. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res.* 38, e197 (2010).

26. Song, N., Liang, A. & Bu, C. A molecular phylogeny of Hemiptera inferred from mitochondrial genome sequences. *PLoS One* 7, e48778 (2012).

27. Cui, Y. et al. Phylogenomics of Hemiptera (Insecta Paraneoptera) based on mitochondrial genomes. *Syst. Entomol.* 38, 233–245 (2013).

28. Ma, C. et al. The compact mitochondrial genome of Zorotypus medoensis, provides insights into phylogenetic position of Zoraptera. *BMC Genomics* 15, 1156 (2014).

29. Coates, B. S. Assembly and annotation of full mitochondrial genomes for the corn rootworm species, *Diabrotica virgifera virgifera* and *Diabrotica barberi* (Insecta: Coleoptera: Chrysomelidae), using next generation sequence data. *Gene* 542, 190–197 (2014).

30. Crampton-Platt, A. et al. Soup to Tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Mol. Biol. Evol.* 32, 2302–2316 (2015).

31. Guillot, C. F. et al. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Mol. Biol. Evol.* 31, 2223–2237 (2014).

32. Rubinstein, N. D. et al. Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastid ascidian mitochondrial genomes. *Genome Biol. Evol.* 5, 1185–1199 (2013).

33. Andújar, C. et al. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Mol. Ecol.* 24, 3603–3617 (2015).

34. Timmermans, M. J. et al. Family-level sampling of mitochondrial genomes in Coleoptera: compositional heterogeneity and phylogenetics. *Genome Biol. Evol.* 8, 161–175 (2016).

35. Song, N., Li, H., Song, F. & Cai, W. Molecular phylogeny of Polyneoptera (Insecta) inferred from expanded mitogenome data. *Sci. Rep.* 6, 36175 (2016).

36. Andrews, S. FastQC: A quality control tool for high throughput sequence data. Bioinformatics Group, Babraham Institute, Cambridge, United Kingdom, (www.bioinformatics.babraham.ac.uk/projects/fastqc) (2010).

37. Andrews, S. FastQC: A quality control tool for high throughput sequence data. Bioinformatics Group, Babraham Institute, Cambridge, United Kingdom, (www.bioinformatics.babraham.ac.uk/projects/fastqc) (2010).

38. Peng, V., Leung, H. C., Liu, S. M. & Chin, F. Y. IDRA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428 (2012).

39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).

40. García-Alcalde, F. et al. Quailmap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28, 2678–2679 (2012).

41. Milne, I. et al. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14, 193–202 (2013).

42. Benet, M. et al. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 69, 313–319 (2013).

43. Bininda-Emonds, O. R. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 6, 156 (2005).

44. Katoch, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).

45. Tamura, K. et al. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729 (2013).

46. Kuck, P. & Meusmann, K. FAScomCAT: Convenient handling of data matrices. *Mol. Phylogenet. Evol.* 56, 1115–1118 (2010).
Author Contributions
N.S. and W.C. designed the research. N.S. performed the experiments. N.S. and H.L. analyzed the data. N.S. wrote the paper. W.C. and H.L. made the insect drawings in Figs 2 and 3. All authors discussed results and implications.

Additional Information
Supplementary information accompanies this paper at doi:10.1038/s41598-017-11132-0

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017