RESEARCH ARTICLE

Open Access

CrossMark

Compositional and mutational rate heterogeneity in mitochondrial genomes and its effect on the phylogenetic inferences of Cimicomorpha (Hemiptera: Heteroptera)

Huanhuan Yang[1], Teng Li[2*], Kai Dang[1] and Wenjun Bu[1*]

Abstract

Background: Mitochondrial genome (mt-genome) data can potentially return artefactual relationships in the higher-level phylogenetic inference of insects due to the biases of accelerated substitution rates and compositional heterogeneity. Previous studies based on mt-genome data alone showed a paraphyly of Cimicomorpha (Insecta, Hemiptera) due to the positions of the families Tingidae and Reduviidae rather than the monophyly that was supported based on morphological characters, morphological and molecular combined data and large scale molecular datasets. Various strategies have been proposed to ameliorate the effects of potential mt-genome biases, including dense taxon sampling, removal of third codon positions or purine-pyrimidine coding and the use of site-heterogeneous models. In this study, we sequenced the mt-genomes of five additional Tingidae species and discussed the compositional and mutational rate heterogeneity in mt-genomes and its effect on the phylogenetic inferences of Cimicomorpha by implementing the bias-reduction strategies mentioned above.

Results: Heterogeneity in nucleotide composition and mutational biases were found in mt protein-coding genes, and the third codon exhibited high levels of saturation. Dense taxon sampling of Tingidae and Reduviidae and the other common strategies mentioned above were insufficient to recover the monophyly of the well-established group Cimicomorpha. When the sites with weak phylogenetic signals in the dataset were removed, the remaining dataset of mt-genomes can support the monophyly of Cimicomorpha; this support demonstrates that mt-genomes possess strong phylogenetic signals for the inference of higher-level phylogeny of this group. Comparison of the ratio of the removal of amino acids for each PCG showed that ATP8 has the highest ratio while CO1 has the lowest. This pattern is largely congruent with the evolutionary rate of 13 PCGs that ATP8 represents the highest evolutionary rate, whereas CO1 appears to be the lowest. Notably, the value of Ka/Ks ratios of all PCGs is less than 1, indicating that these genes are likely evolving under purifying selection.

Conclusions: Our results demonstrate that mt-genomes have sites with strong phylogenetic signals for the inference of higher-level phylogeny of Cimicomorpha. Consequently, bioinformatic approaches to removing sites with weak phylogenetic signals in mt-genome without relying on an a priori tree topology would greatly improve this field.

Keywords: Mitochondrial genomes, Heterogeneity, Cimicomorpha, Tingidae, Reduviidae, Phylogeny

* Correspondence: teng.chn@gmail.com; wenjunbu@nankai.edu.cn
[2]Institute of Zoology and Developmental Biology, College of Life Sciences, Lanzhou University, 222 Tianshui South Road, Lanzhou 730000, China
[1]Institute of Entomology, College of Life Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China

Yang *et al. BMC Genomics* (2018) 19:264

Page 2 of 13

Background

The number of complete or nearly complete mitochondrial genomes (mt-genomes) of insects has rapidly increased in recent years with the development of next-generation sequencing technologies [1]. Although widely utilized in phylogenetic analyses, potential biases such as high percentages of AT content, base-compositional heterogeneity between lineages, and high evolutionary rates have all been documented in insect mt-genomes [1–4]. These anomalous characteristics frequently limit their applicability in higher-level phylogenetic reconstruction of insects, leading to an incongruence with morphological and nuclear data [1–4].

These potential biases may lead to systematic errors when the evolutionary model used for phylogenetic inference does not take them into account [3]. Compositional heterogeneity across lineages and sequence saturation due to accelerated substitution rates are two common phenomenon causing phylogenetic artefacts in mt-genome data [5, 6]. The phenomena caused by these potential biases have attracted an increasing amount of attention in studies of some groups, such as Coleoptera, Thysanoptera, Psocodea, Sternorrhyncha (Hemiptera), Strepsiptera, Neuropterida and Hymenoptera [1–3, 7–18], showing that compositional heterogeneity is pervasive. If these characteristics were shared by unrelated lineages, the apparent convergent evolution may erode a genuine phylogenetic signal [6]. Moreover, accelerated evolutionary rates and rate variation across lineages can increase susceptibility to systematic errors, such as long-branch attraction (LBA) [19–21].

Various strategies have been proposed to ameliorate the effects of such biases. For example, first, dense taxon sampling may contribute to inferring multiple substitutions at a site correctly and result in improvements in the estimation of tree topology through improving the estimation of molecular rates and variation in base composition [5, 10, 11, 13, 22]. In addition, the strategy of sampling more taxa to break up long branches has been widely advised to decrease the effects of LBA bias [10, 23–25]. Second, removing the third codon positions from protein-coding genes (PCGs) or using purine-pyrimidine (RY) coding can reduce the effects of compositional heterogeneity and saturation in the assessment of character variation [10, 11, 13, 26–29]. Finally, using evolutionary models such as the site-heterogeneous mixture model [30], an alternative strategy for accommodating complex character variation might mitigate the impact of compositional and mutational bias [3, 10–13, 17, 31–33]. In fact, it has already been shown that the site-heterogeneous mixture model CAT was able to overcome LBA artefacts in some groups [3, 19, 34–37].

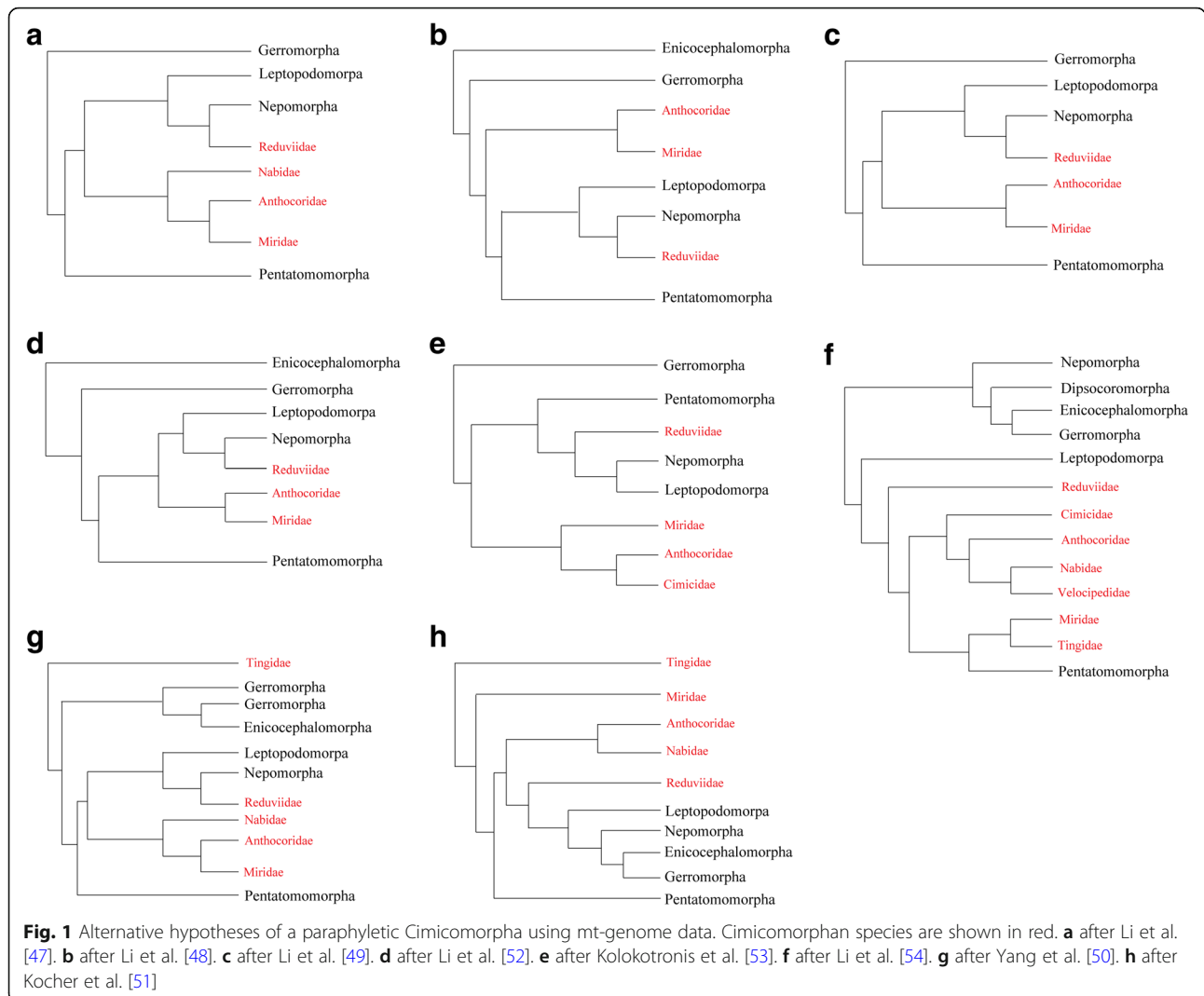Cimicomorpha is the largest infraorder in Heteroptera, with 17 families and more than 20,000 species [38]. The monophyly of Cimicomorpha has been consistently and strongly supported by most previous studies based on morphological characters [39–41], molecular data [42–44], and combined data analyses [45], and recently, it was also supported based on the transcriptomic data of 53 taxa and 3102 orthologous genes [46]. However, Cimicomorpha was recovered as a paraphyletic group when the data for mt-genomes alone were used in the analyses [47–54]. The paraphyly of Cimicomorpha was caused mainly by the positions of the families of Reduviidae and Tingidae (Fig. 1). Reduviidae, known as assassin bugs, most of which are predatory insects, is a diverse group of approximately 7000 described species worldwide [55–57]. When the mt-genomes of Reduviidae species were added to the dataset, Reduviidae separated from Cimicomorpha and became the sister group to Nepomorpha (Fig. 1a, b, c, d) or Nepomorpha + Leptopodomorpha (Fig. 1e). In addition, recent studies conducted by Li et al. [54] showed that Reduviidae was supported as the sister group of a clade that included Pentatomomorpha and the remainder of Cimicomorpha (Fig. 1f). Tingidae, known as lace bugs, is a family of approximately 2500 species [58]. When the mt-genomes of a few Tingidae species were used, Tingidae became the sister group to a clade that consisted of the remainder of Heteroptera (Fig. 1g, h), rather than a member of Cimicomorpha. In the above phylogenetic studies recovering a paraphyletic Cimicomorpha, the PCG third codon positions were not removed [47–51, 53], and the analyses were not inferred using heterogeneous models, except in the study by Li et al. [52, 54]; thus, we attempted to incorporate these approaches.

Currently, the mt-genomes of 13 of 18 species of Reduviidae (which belong to 8 subfamilies, 13 genus) reported to date were used in our analyses, while only 4 complete mt-genomes belonging to Tinginae have been sequenced for lace bugs. Here, we sequenced five additional Tingidae species and attempted to discuss the compositional and mutational rate heterogeneity in mt-genomes and its effect on the phylogenetic inferences of Cimicomorpha incorporating the bias-reduction strategies described above.

Results

General features of mt-genomes of five newly sequenced species of Tingidae

In this study, we sequenced the mt-genomes of five species of Tingidae (Additional file 1), including one complete (*Trachypeplus jacobsoni*) and four nearly complete (*Cysteochila chiniana*, *Dictyla platyoma*, *Metasalis populi*, and *Tingis cardui*) mt-genomes. In these nearly complete mt-genomes, the control region is incomplete due to highly repetitive regions that could not be assembled with

Yang *et al. BMC Genomics* (2018) 19:264

Page 3 of 13

Fig. 1 Alternative hypotheses of a paraphyletic Cimicomorpha using mt-genome data. Cimicomorphan species are shown in red. **a** after Li et al. [47]. **b** after Li et al. [48]. **c** after Li et al. [49]. **d** after Li et al. [52]. **e** after Kolokotronis et al. [53]. **f** after Li et al. [54]. **g** after Yang et al. [50]. **h** after Kocher et al. [51]
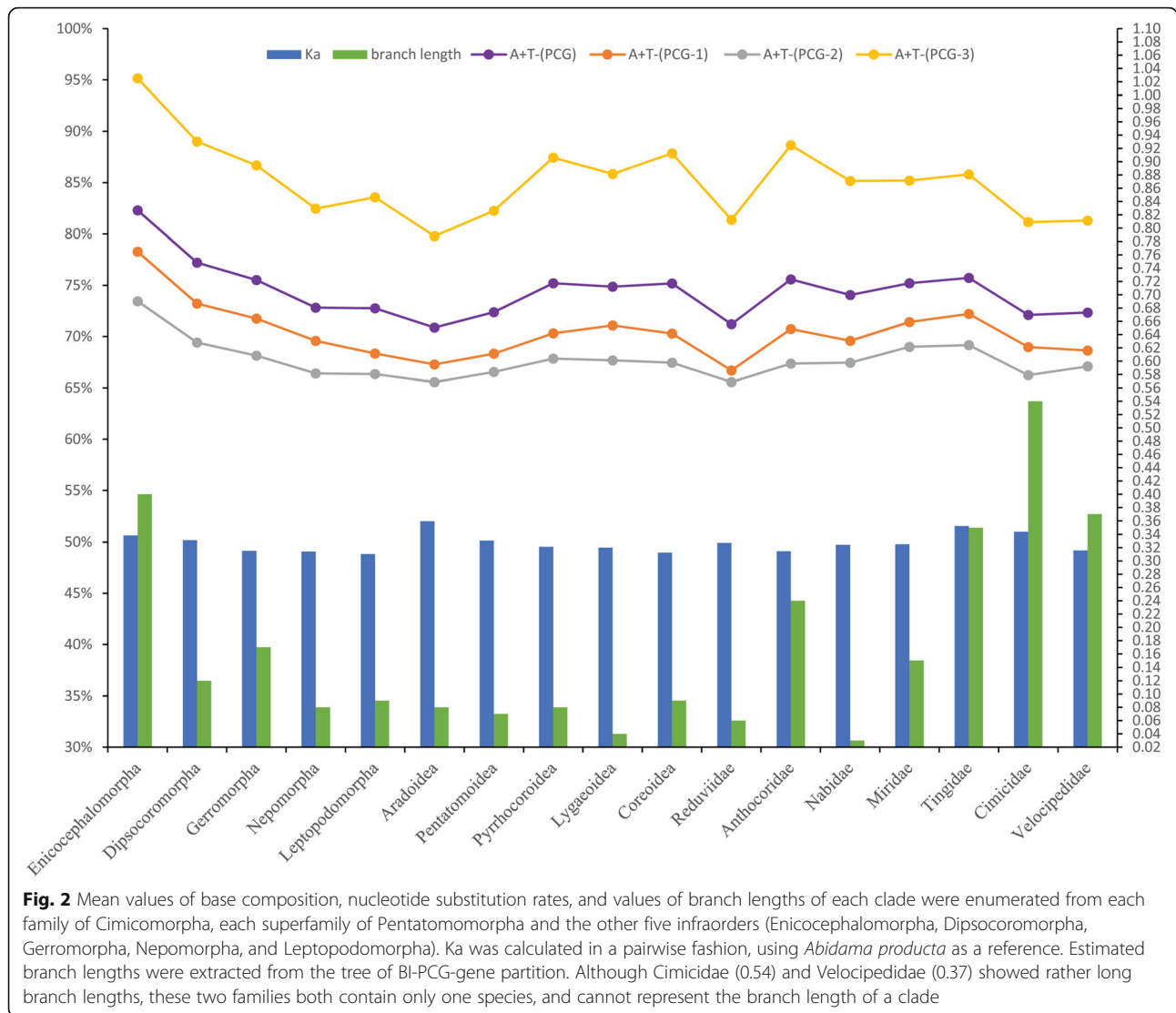
certainty. Each mt-genome contained the typical 37 genes (2 rRNAs, 13 PCGs and 22 tRNAs) with the same gene order observed in the four previously sequenced lace bug mt-genomes [50, 51, 54] and found in most other insects [1]. In the five newly sequenced Tingidae mt-genomes, all PCGs initiated with ATN as the start codon, except for the ATP6 gene, which was only found in *Cysteochila chiniana* and started with GTG. Most PCGs ended with the termination codons TAA/TAG, whereas the remaining PCGs were terminated with T. All typical 22 animal tRNA and 2 rRNA genes were observed in each of these mt-genomes.

High degree of compositional heterogeneity

The AT content of PCGs from the included heteropteran species ranged from 65.8% (*Reduvius tenebrosus*) to 83. 2% (*Stenopirates* sp.) with a mean of 74.0%. Tingidae PCGs had AT contents between 72.8% (*Cysteochila*

chiniana) and 78.2% (*Stephanitis mendica*) with a mean of 75.7% (Fig. 2), which is the highest family-level mean value within the true bugs, except for a few families represented by only one or two taxa. In contrast, the AT content of Reduviidae was between 65.8% (*Reduvius tenebrosus*) and 74.4% (*Oncocephalus breviscutum*) with a mean of 71.2%, which is the lowest family mean of the heteropterans. Analyses of base composition at each codon position across heteropteran insects showed that third codon positions were much higher in AT content than first and second codon positions (Fig. 2).

Individual chi-squared tests of each PCG indicated that all genes were heterogeneous ($P < 0.05$ that the data represented compositional heterogeneity) (Additional file 2). The third codon positions showed significant heterogeneity for all genes ($P = 0$), while all second codon positions appeared homogeneous. The first codon positions revealed heterogeneity ($P < 0.05$) in the genes ND2, ND4,

Yang *et al. BMC Genomics* (2018) 19:264

Page 4 of 13

Fig. 2 Mean values of base composition, nucleotide substitution rates, and values of branch lengths of each clade were enumerated from each family of Cimicomorpha, each superfamily of Pentatomomorpha and the other five infraorders (Enicocephalomorpha, Dipsocoromorpha, Gerromorpha, Nepomorpha, and Leptopodomorpha). Ka was calculated in a pairwise fashion, using *Abidama producta* as a reference. Estimated branch lengths were extracted from the tree of BI-PCG-gene partition. Although Cimicidae (0.54) and Velocipedidae (0.37) showed rather long branch lengths, these two families both contain only one species, and cannot represent the branch length of a clade

ND5 and ND6, but other PCGs showed compositional homogeneity. When the first and third codon positions were RY recoded and each gene was reanalysed, the first codon positions were no longer apparently heterogeneous in any gene except ND2, but the third positions of most PCGs still remained compositionally heterogeneous in seven genes, including ATP6, CO1, CytB, ND1, ND2, ND4 and ND5 (Additional file 2).

In addition, saturation analyses of each codon position of all PCGs indicated substantial saturation of the substitutions in the third codon positions, while the first and second codon positions were free of saturation (Additional file 3). The AliGROOVE procedure [59] was employed to detect heterogeneous sequence divergence, demonstrating that third codon positions were much more rate-heterogeneous than first and second positions and consistently scored negative in pairwise comparisons (Additional file 4).

Contrasting rates of evolution in specific lineages

We calculated Ka (non-synonymous substitution rate) for each species in this study using *Abidama producta* (Cercopidae: Auchenorrhyncha) as a reference. In Heteroptera, these comparisons showed that Ka ranged from 0. 30 (*Stictopleurus subviridis*) to 0.37 (*Neuroctenus parus*) with a mean of 0.33 (Fig. 2). Notably, Tingidae (0.34–0.36) had the highest family-level mean Ka (0.35) of true bugs, except Aradidae (0.36), which contained only two taxa, while the mean value of Ka (0.33) in Reduviidae was comparable to that of heteropteran insects as a whole. A comparison of branch lengths (Fig. 2) in the phylogenetic tree revealed that Tingidae (0.35) exhibited much longer branch length than the other Heteropteran species except the Enicocephalomorpha (0.40), while Ruduviidae (0.06) had the shortest branch length among Heteropteran species except the Nabidae (0.03) and Lygaeoidea (0.04). Overall, these results indicated

Yang *et al. BMC Genomics* (2018) 19:264

Page 5 of 13

contrasting substitution rates among different heteropteran lineages.

Phylogenetic analyses

Dense taxa sampling of Tingidae and Reduviidae

We performed phylogenetic analyses on the mt-genomes of Cimicomorpha with a dense taxon sampling of Tingidae and Reduviidae. Maximum Likelihood (ML) and Bayesian Inference (BI) analyses under homogeneous models indicated that Tingidae was sister to the remaining Heteroptera (Fig. 3), as reported by previous studies based on mt-genomes [50, 51]. A paraphyletic Cimicomorpha was split into three clades Miridae + Cimicidae + Anthocoridae + Velocipedidae + Nabidae, Reduviidae (sister to remaining heteropteran infraorders except Pentatomomorpha and the remainder Cimicomorpha), and Tingidae (sister to remaining heteroptera), which is incongruent with previous studies based on morphological characteristics [39–41] and/or molecular data [42–46]. In addition, BI and ML analyses of an optimized partition scheme (PCG-gene partition and PCG-codon partition) also found a paraphyletic Cimicomorpha (Fig. 3). All these analyses indicated that dense taxon sampling alone did

not resolve the problem of artefactual paraphyly of Cimicomorpha.

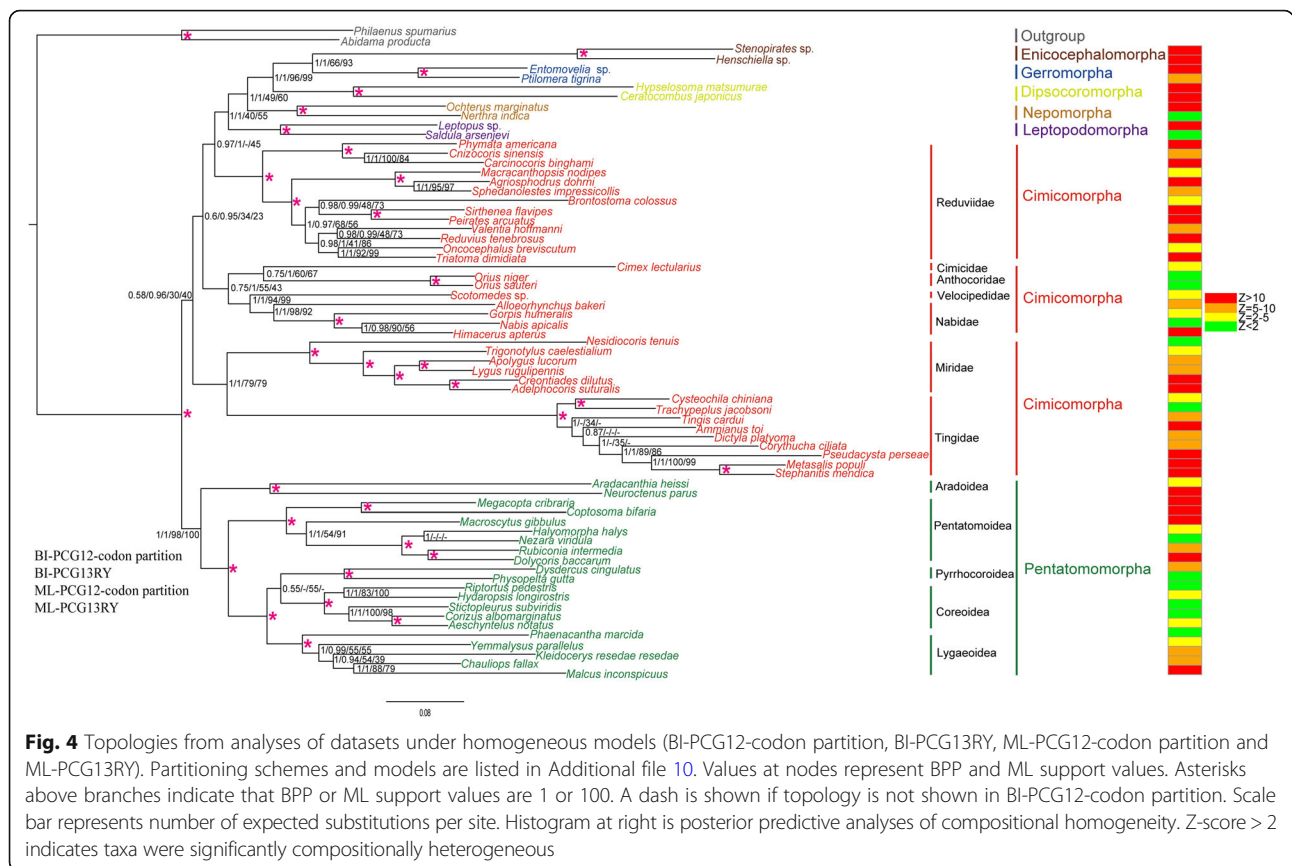### Removal of third codon positions of PCGs or RY coding

We excluded the third codon position of PCGs or used RY coding of the first and third codon positions to reduce the effects of compositional heterogeneity. In both BI and ML analyses, removing the third codon positions or used RY coding of first and third codon positions resulted in the monophyly of Miroidea (Tingidae + Miridae) (Fig. 4). The paraphyly of Cimicomorpha changed to three different clades: Reduviidae (sister to the rest of heteroptera, except Pentatomomorpha and the remaining Cimicomorpha), Cimicidae + Anthocoridae + Velocipedidae + Nabidae (sister to remaining heteropteran species except Miroidea and Pentatomomorpha) and Tingidae + Miridae (sister to remaining heteropteran infraorders, except Pentatomomorpha). The monophyly of Reduviidae, Cimicidae, Anthocoridae, Velocipedidae, Nabidae, Tingidae and Miridae were all strongly supported by both Bayesian posterior probabilities (BPP) and bootstrap (BS) values.

Comparing saturation plot slopes estimated by plotting observed distances (uncorrected P-distances) against patristic distances from the CAT + GTR model indicated
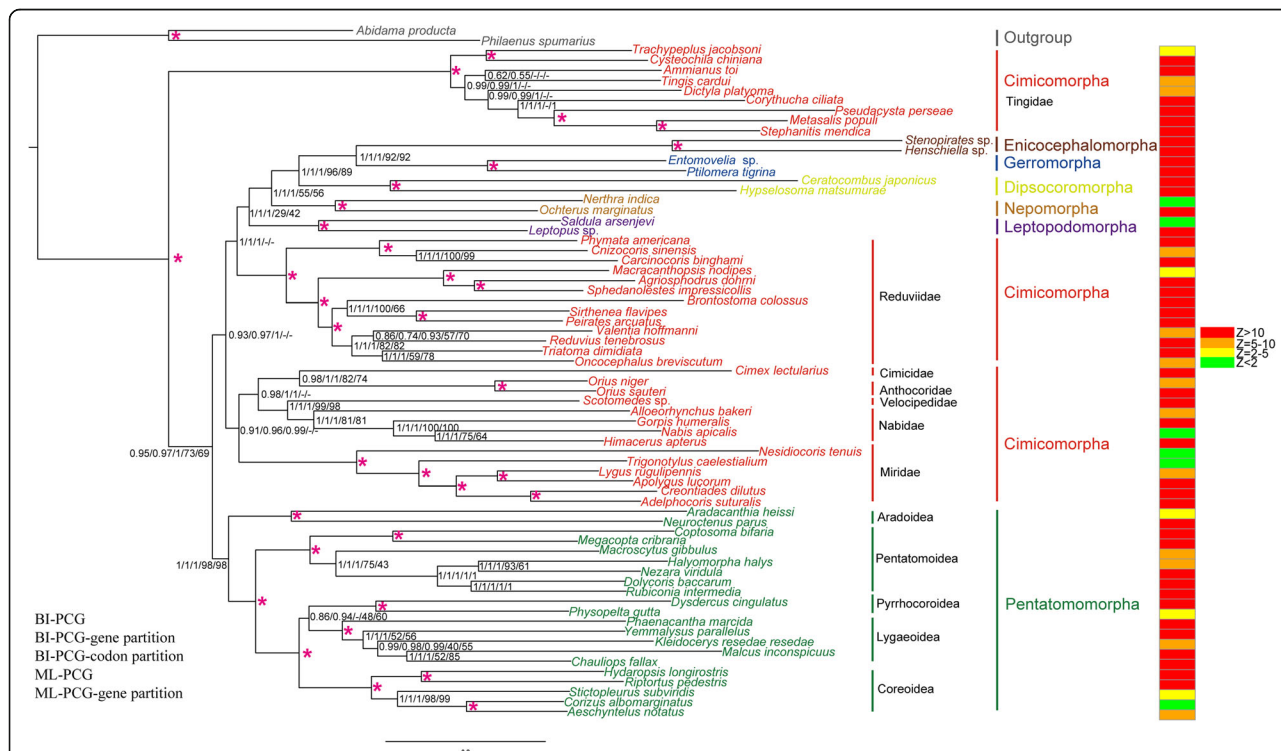
Fig. 3 Topologies from analyses of datasets under homogeneous models (BI-PCG, BI-PCG gene partition, BI-PCG-codon partition, ML-PCG, ML-PCG-gene partition). Partitioning schemes and models are listed in Additional file 10. Values at nodes represent Bayesian posterior probabilities (BPP) and ML support values. Asterisks above branches indicate that BPP or ML support values are 1 or 100. A dash is shown if topology is not shown in BI- PCG. Scale bar represents number of expected substitutions per site. Histogram at right is posterior predictive analyses of compositional homogeneity. Z-score > 2 indicated taxa were significantly compositionally heterogeneous

Yang *et al. BMC Genomics* (2018) 19:264

Page 6 of 13

Fig. 4 Topologies from analyses of datasets under homogeneous models (BI-PCG12-codon partition, BI-PCG13RY, ML-PCG12-codon partition and ML-PCG13RY). Partitioning schemes and models are listed in Additional file 10. Values at nodes represent BPP and ML support values. Asterisks above branches indicate that BPP or ML support values are 1 or 100. A dash is shown if topology is not shown in BI-PCG12-codon partition. Scale bar represents number of expected substitutions per site. Histogram at right is posterior predictive analyses of compositional homogeneity. Z-score > 2 indicates taxa were significantly compositionally heterogeneous

that the dataset PCG12 showed the lowest level of saturation (PCG12 = 0.0119, PCG = 0.0022 and AA = 0.0104) (Additional file 5).

Under heterogeneous models

The major changes that resulted from using the CAT + GTR model with three datasets (PCG, PCGRNA and AA) (Fig. 5) were that, although the monophyly of Cimicomorpha was still not recovered, all Cimicomorpha and Pentatomomorpha form a monophyletic group. Cimicomorpha was split into two groups: (Cimicidae + Anthocoridae + Velocipedidae + Nabidae + Tingidae + Miridae) and Reduviidae (sister group of a clade that included Pentatomomorpha and the remainder of Cimicomorpha) (Fig. 5). Moreover, GTR models inferred a much lower level of homoplasy in the nucleotide dataset (PCG, PCG12 and PCGexclude) compared to the CAT + GTR model in posterior predictive analysis (Additional file 6).

Removal of sites with weak phylogenetic signals of mt-genomes

Based on the tree topology of Wang et al. [44] in supporting the monophyly of Cimicomorpha, and consistent with most previous studies based on morphological characters, the morphological and molecular

combined data and large scale molecular datasets, 4909 sites (497 complete amino acids) with weak phylogenetic signals of PCG dataset were excluded to generate the dataset PCGexclude under a strict filter criterion. The dataset PCGexclude can support the monophyly of Cimicomorpha, which demonstrates that mt-genomes possess strong phylogenetic signals for the inference of higher-level phylogeny of this group. Posterior predictive analyses of compositional homogeneity in PCGexclude indicated that it had the lowest level of heterogeneity, with 45 among 67 species that were still compositionally heterogeneous (Additional file 7), compared to 61 species in PCG (Fig. 3) and 54 species in PCG12 (Fig. 4). The results indicated that removal of sites with weak phylogenetic signals clearly reduced the degrees of heterogeneity.

A comparison of the ratio of the excluded amino acids/all 3709 amino acids for each PCG showed that ATP8 has the highest ratio of removed amino acids, while CO1 has the least (Fig. 6). This pattern is largely congruent with the evolutionary rate of 13 PCGs that ATP8 represents the highest evolutionary rate, whereas CO1 appears to be the lowest. The evolutionary patterns of 13 PCGs in our study were consistent with previous studies [48, 52, 60].

Yang *et al. BMC Genomics* (2018) 19:264

Page 7 of 13

Fig. 5 Phylogeny inferred from PhyloBayes analyses under CAT+GTR model with datasets PCGRNA, PCG, and AA. Values at nodes represent BPP. Asterisks above branches indicate support values of 1. A dash is shown if the topology is not shown in PCGRNA. Scale bar represents number of expected substitutions per site

Fig. 6 Ratio of excluded amino acids and evolutionary rates of 13 PCGs in Heteroptera. Ratio of excluded amino acids/all amino acids, rate of non-synonymous substitutions (Ka), rate of synonymous substitutions (Ks) and ratio of rate of non-synonymous substitutions to rate of synonymous substitutions (Ka/Ks) were calculated for each PCG

Yang *et al. BMC Genomics* (2018) 19:264

Page 8 of 13

Discussion

The phylogenetic position of Tingidae and Reduviidae

The monophyly of Cimicomorpha has been consistently supported by studies based on morphological characteristics [39–41], molecular data [42–44] and combined data analyses [45]. Recently, it has also been supported by transcriptomic data of 53 taxa and 3102 orthologous genes [46]. In contrast, phylogenetic analyses based on concatenated sequences of mt-genes always failed to recover Cimicomorpha as a monophyletic group [47–54], caused mainly by the positions of the families of Reduviidae and Tingidae, especially Tingidae, which was placed as the sister group to all remaining heteropterans by Yang et al. [50] and Kocher et al. [51]. These surprising phylogenetic results have no support from morphological data or nuclear gene sequences. We hypothesized that the findings of Yang et al. [50] and Kocher et al. [51] might be due to biases introduced by inadequate taxon sampling, as only one or two Tingidae mt-genomes were included. Increased taxa sampling can often lead to more accurate phylogenetic inference [5, 10, 11, 13, 22], but increased taxa sampling for Tingidae and Reduviidae in our analyses did not obviously improve the result under homogeneous models. Even using data partitioning, the effect was still limited: Tingidae remained the sister-group of a clade that included the remaining species of Heteroptera (Fig. 3).

Because the dense taxon sampling of Tingidae and Reduviidae in our analyses could not solve this thorny problem, we examined the possibility that these findings resulted from base compositional heterogeneity and accelerated evolutionary rates of mt-genomes implied using inappropriate models [6, 7, 17]. Our analyses under homogeneous models indicated that after the removal of the third codon positions, the Tingidae was recovered as the sister group of Miridae entirely outside of all Cimicomorpha, as previously reported (Fig. 4). RY coding is generally thought to decrease saturation and compositional bias [10, 27], but when RY recoding was used in our analyses, Cimicomorpha was still a paraphyletic group (Fig. 4). After verifying the compositional and mutational heterogeneity of the third codon positions, we suspect that the artefactual phylogenetic result might be due to the use of inappropriate models.

Sophisticated evolutionary models, such as the heterogeneous CAT + GTR model, which account for among-site heterogeneity, can lessen the effects of compositional biases and better reflect the evolutionary process [10, 11, 30]. The use of a heterogeneous model predicted homoplasies more efficiently than homogeneous models in our dataset (Additional file 6). Bayesian analyses using PhyloBayes MPI Version 1.7 [61] for the amino acid and nucleotide datasets using heterogeneous CAT + GTR recovered the monophyly of (Tingidae + Miridae). Although the monophyly of Cimicomorpha was not recovered, as (Reduviidae + (other Cimicomorpha + Pentatomomorpha)) formed a clade, it indeed showed a significant improvement over previous studies [47–54] and recovered the monophyly of Terheteroptera (Cimicomorpha and Pentatomomorpha) based on mt-genome data. Our results confirmed the power of site-heterogeneous mixture models for resolving phylogenetic relationships with Cimicomorpha and showed the significance of adequate model selection. This significance suggests that site-heterogeneous models may be preferable models for phylogenetic reconstruction when using mt-genomes.

Several commonly suggested strategies to reduce sources of systematic bias were used in the phylogenetic inference of Cimicomorpha. Our results demonstrate that these strategies (local dense taxon sampling, removal of third codon positions, RY coding and use of the site-heterogeneous model) were insufficient to recover the monophyly of the Cimicomorpha. The phylogenetic relationships based on the dataset removed the weak phylogenetic sites that were consistent with the results based on large scale dataset (Additional file 7), which indicated that the mt-genomes possess strong phylogenetic signals for the inference of higher-level phylogeny of Cimicomorpha. While these sites with weak phylogenetic signals are only used for the inference of higher-level phylogeny, they may still have strong phylogenetic signals for the inference of lower-level phylogeny, such as the genus or species levels.

Effect of compositional and mutational rate heterogeneity

A high degree of compositional heterogeneity across all PCGs was found in our dataset, with the third codon position showing significant levels of compositional heterogeneity and to a lesser extent the first and second positions, even after RY coding (Additional file 2). Moreover, ND2, 4, 5, and 6 genes are more compositionally heterogeneous than the other PCGs (Additional file 2), and this phenomenon was also observed in a detailed analysis of Coleoptera [10]. This phenomenon-that heterogeneity mainly affected the NADH genes, which are associated with functionality-might indicate that the compositional heterogeneity of the NADH genes is driven by variation in the protein level and possible co-variation in the NADH protein complex [10].

Accelerated substitution rates may also play a role in eroding phylogenetic signal through unrecognized homoplasy and can lead to problems with LBA [25, 62, 63]. Species of the Aradidae and Tingidae genera had a markedly accelerated evolutionary rate (Fig. 2). Generally, both species of the two families are weakly flying

Yang *et al. BMC Genomics* (2018) 19:264

Page 9 of 13

species. Mitochondria, via oxidative phosphorylation, supply most of the energy required for locomotion, and the metabolic power required for locomotion has a linear correlation with speed [64]; we speculate that with the degeneration of locomotive ability, weakly flying or flightless species might require less energy efficient metabolism than rapidly flying taxa. Thus, the functional constraints are relaxed on the mt-genomes, which might accumulate more nucleotide substitutions [65].

Removal of sites with weak phylogenetic signals

In our analyses, 4909 sites (including 497 complete amino acids) with weak phylogenetic signals were excluded from 11,127 sites (3709 amino acids) of PCGs. The comparison of the ratio of the excluded amino acids/all 3709 amino acids for each PCG showed that ATP8 has the most amino acids removed, while CO1 has the least; this finding is consistent with the pattern of the evolutionary rate of 13 PCGs, in which ATP8 represents the highest evolutionary rate and CO1 appears to be the lowest (Fig. 6). Notably, the value of Ka/Ks ratios of all PCGs are less than 1, indicating that these genes are likely evolving under purifying selection [52, 66]. In the evolution of mt-genomes, purifying selection is the predominant force [67, 68]. It is possible that when lifestyle changes with greater energy demands or reduced oxygen availability, in this context of strong purifying selection, weak and/or episodic positive selection occurs [67, 68]. The evolutionary patterns of 13 PCGs in our study were consistent with previous studies [48, 52, 60]. As the proteins encoded by mt-genomes provide up to 95% of the cell's energy requirements, they play a critical role in oxidative phosphorylation [65]. Nonsynonymous substitutions can cause defects in respiratory-chain activity that reduce the efficiency of metabolic processes and are generally more harmful [69, 70]. To maintain functional requirements, the amino acids of CO1 experienced strong evolutionary constraints [71, 72] and undergo strong evolutionary pressures. While ATP8 underwent weak evolutionary pressures and functional constraints, this relaxation of metabolic constraint may enable the accumulation of more mutations in the mt-genomes.

Conclusions

Our results indicate that it is a challenge to reconstruct the phylogenetic relationships of Heteroptera based on mt-genomes, especially for Cimicomorpha, due to its high degree of compositional heterogeneity and significantly accelerated evolutionary rates in specific lineages. When the strategies of the appropriate model were selected, such as site-heterogeneous mixture models, the monophyly of Terheteroptera (Cimicomorpha and Pentatomomorpha) based on mt-genome data was recovered. Unfortunately, these suggestions cannot completely remove the potential biases in this group, complicating the inference of higher-level phylogeny of Cimicomorpha.

The removal of sites with weak phylogenetic signals from the mt-genomes dataset based on a parsimony-based approach—which strongly relies on a previous generally accepted tree topology—can reduce these potential biases significantly. The phylogenetic relationships of Cimicomorpha based on the dataset, which removed the sites with weak phylogenetic signals, were consistent with the results based on various datasets in previous studies and demonstrated that mt-genomes possess strong phylogenetic signals for the inference of higher-level phylogeny. This analysis can measure how much noisy data must be removed to reduce the impact of weak phylogenetic signal, and we can also determine the source of the noise sites, including particular genes and particular amino-acid residues. The problem will be how to identify sites with weak phylogenetic signals in mt-genome datasets without relying on the topology of an a priori tree. Consequently, bioinformaticians will need to propose new approaches without an a priori tree to uncover these sites in mt-genome datasets, which will greatly improve this field.

Methods

Taxa sampling and sequencing

In total, 67 species of Heteroptera, representing all seven infraorders of Heteroptera, were included, with two species of Auchenorrhyncha (*Philaenus spumarius* and *Abidama producta*) as outgroups, as Cicadomorpha was strongly supported as the sister group of Heteroptera in a previous study based on mt-genomes [73]. The mt-genomes of five lace bug species, *Trachypeplus jacobsoni*, *Cysteochila chiniana*, *Dictyla platyoma*, *Metasalis populi*, and *Tingis cardui,* are reported here for the first time. The remainder were retrieved from GenBank, including 24 mt-genomes published previously by our group [25, 44, 49, 74]. Detailed taxon information is included in Additional files 8 and 9.

For the newly sequenced species, total genomic DNA was extracted from thoracic muscle tissue using a CTAB-based method [75]. Then, the entire mt-genome was obtained using the Illumina HiSeq 2000 platform (Illumina, San Diego, CA) with a 200-bp insert size and a paired-end 100-bp sequencing strategy at BGI-Shenzhen, China. The identification of PCGs and ribosomal RNA genes (rRNAs) was performed as in previous studies [52]. The annotation of five newly sequenced mt-genomes of Tingidae is provided in Additional file 1.

Sequence alignment and dataset concatenation

The sequences of 13 PCGs and 2 rRNAs from mt-genomes were used in our analyses. The methods for the

Yang *et al. BMC Genomics* (2018) 19:264

Page 10 of 13

alignment of PCGs and rRNAs were same as in previous studies [52]. Alignments of individual genes were concatenated to generate various datasets to reconstruct the phylogeny: 1) PCG: all three codon positions of the 13 PCGs; 2) PCGRNA: all three codon positions of the 13 PCGs and two rRNAs; 3) AA: amino acid sequences of 13 PCGs. Furthermore, we used Partition-Finder 2.0 [76] to test the various partitioning schemes for ML and BI methods, and the input configuration files, which contained different predefined partitions for each dataset, were created: 1) PCG-gene partition: 13 gene partitions for PCGs; 2) PCG-codon partition: 39 codon partitions for PCGs; 3) PCG12-codon partition: 26 codon positions for the first and second codon positions. We used the Bayesian information criterion (BIC) and the "greedy" algorithm with branch lengths estimated as "unlinked" to search for the best-fit scheme and substitution model (Additional file [10]). To decrease saturation and compositional bias, we also used RY coding [26, 27, 29] datasets, as PCG13-RY (PCGs with the first and third codon positions RY coded) was used.

Compositional heterogeneous and contrasting rates analyses

Base composition was analysed with MEGA 6.0 [77]. We used the chi-squared statistic to test the compositional heterogeneity of PCGs as described in Foster (2004) [78]. Analysis of heterogeneity was conducted in PAUP*4 [79], including the ingroup only. Based on tail area probabilities (Pt), < 0.05 was deemed to indicate compositional heterogeneity. To test whether the taxa in our dataset are compositionally heterogeneous, we conducted posterior predictive analysis under the CAT +GTR model using PhyloBayesv4.1c [31]. A Z-score value of more than 2 indicated that the taxa were significantly compositionally heterogeneous. In addition, AliGROOVE [59] was used to analyse sequence divergence heterogeneity with the default sliding window size. Ambiguity was set for the nucleotide dataset to generate profiles of pairwise sequence similarity for all pairwise sequence comparisons. To test for substitution saturation, we plotted each codon position based on the K80 model for transition and transversion substitutions in DAMBE V4.5.32 [80].

The rate of non-synonymous substitutions (Ka), synonymous substitutions (Ks) and the ratio of Ka/Ks were calculated for each PCG in DnaSP 5.0 [81]. Estimated branch lengths were extracted from the tree BI-PCG-gene partition. Branch lengths of each clade, i.e., each family of Cimicomorpha, each superfamily of Pentatomomorpha and the five other infraorders (Enicocephalomorpha, Dipsocoromorpha, Gerromorpha, Nepomorpha, and Leptopodomorpha) were enumerated.

As Cimicidae and Velocipedidae both contain only one species, they cannot represent the branch length of a clade.

Model-based saturation plots and posterior predictive analyses

To measure how well the models anticipated sequence saturation and homoplasy, we conducted saturation plots and posterior predictive analyses. The best-fit CAT +GTR model was selected as a reference for saturation plots. Observed distances (uncorrected P-distances) and patristic distances generated by PATRISTIC [82] were then calculated, and the regression from the ordered pairs of distances plotted against each other. The slope of the regression line, indicates the level of saturation: the shallower the slope, the greater the level of saturation. PhyloBayes v4.1c [31] was used to conduct posterior predictive analyses to compare alternative models' ability to estimate homoplasy in our datasets.

Phylogenetic analyses
Using homogeneous models
Phylogenetic analyses were initially conducted using standard BI and ML analyses with homogeneous models. The datasets were not partitioned, and Modeltest 3.7 [83] was used to infer the best substitution models for nucleotide data. BI analyses were conducted using GPU MrBayes [84]. In BI, the GTR + I + G substitution model for nucleotide data was used. Two simultaneous runs with four chains of 10,000,000 generations were conducted for the matrix. Each set was sampled every 1000 generations with a burn-in of 25%. The two runs converged satisfactorily, with a standard deviation of split frequency lower than 0.01, and the effective sample size (ESS) was above 200. ML analyses were conducted using RAxML 8.0.12 [85] with the GTR + I + G model for nucleotide data. Nodal support was calculated with bootstrap values from heuristic searches of 1000 resampled datasets using the rapid bootstrap feature (random seed value 12,345) [86].

Using heterogeneous models
For both AA and nucleotide datasets, PhyloBayes MPI Version 1.7 [61] was used to conduct phylogenetic analyses with the CAT+GTR model. Two independent searches were run until the likelihoods stabilized and the two runs had satisfactorily converged (maxdiff less than 0.3). The initial trees of each run were discarded as burn-in, and a consensus tree was computed from the remaining trees.

Removing sites with weak phylogenetic signals of mt-genomes
To test whether the dataset of mt-genomes have sites with strong phylogenetic signals for phylogenetic inference at higher-level category of Cimicomorpha, we adopted a

Yang *et al. BMC Genomics* (2018) 19:264

Page 11 of 13

parsimony-based approach to detect the weak phylogenetic signals of mt-genomes [87]. The PCG source dataset and the corresponding tree topology from Wang et al. [44] were analysed in PAUP*4.0b10 [79], using DELTRAN optimization. All sequences of the dataset were depicted on the preset branches of the lineages. A labelled tree with a complete list of sites was obtained by activating the log-file options (Describetrees/root = outgroup, plot = phylogram, labelnode = yes, apolist = yes). The consistency index (CI) of each site was used as a filter criterion. It is generally recognized that when the CI value of each site was lower than 0.3, this site was likely probably caused by homoplasy, which means with weak phylogenetic signal [88]. We selected CI (0.3) to filter the sites obtained in the parsimony-based analyses through a series of in-house shell scripts (Additional file 11). Finally, the generated sub-dataset (PCGexclude) excluded the sites with CI values below 0.3 for subsequent analyses.

Additional files

Additional file 1: Annotation and organization of five Tingidae mt-genomes sequenced in this study. (XLSX 19 kb)

Additional file 2: Conventional chi-squared test of each gene and dataset with each codon position. $P < 0.05$ indicated heterogeneity. PCG1, the first codon position of PCG. PCG2, the second codon position of PCG. PCG3, the third codon position of PCG. PCGRY1, the first codon position was RY recoded. PCGRY3, the third codon position was RY recoded. (TIFF 502 kb)

Additional file 3: Substitution patterns of all codon positions. The number of transition (S) and transversion (V) substitutions are plotted against Kimura 2-parameter (K2p) distance, considering all sites. Each point represents pairwise comparison among two taxa. (TIFF 730 kb)

Additional file 4: Heterogeneous sequence divergence within heteropteran mt-genomes. The mean similarity score between sequences was represented by a coloured square. Scores range from − 1 (indicating a maximally random level of similarity), to + 1 (indicating maximally non-random similarity). Darker red indicates more randomized accordance between the pairwise sequence comparisons. Blue indicates a less randomized accordance. The dataset name of each codon position is listed on the bottom left corner. (TIFF 2403 kb)

Additional file 5: Model-based saturation plots for amino acid and nucleotide datasets. Plots of patristic distances of datasets (PCG, AA and PCG12) as estimated from the CAT+GTR tree, compared to distances estimated from the observed distances (uncorrected P-distances). (TIFF 459 kb)

Additional file 6: Posterior predictive analyses of sequence homoplasy. DH is the difference between the observed and predicted homoplasy. Values closer to zero indicate better predictions. (XLSX 10 kb)

Additional file 7: Topology based on analyses of dataset PCGexclude under homogeneous models. We show a schematic version of the phylogenetic trees, with some lineages collapsed for clarity. Values at nodes represent BPP and ML support values. Asterisks above the branches indicate that BPP or ML support values are 1 or 100. The scale bar represents the number of expected substitutions per site. The histogram on the right was the posterior predictive analyses of compositional homogeneity. A Z-score > 2 indicated taxa were significantly compositionally heterogeneous. (TIFF 1528 kb)

Additional file 8: Collection information of 5 species newly sequenced in present study. (XLSX 9 kb)

Additional file 9: Complete or nearly complete mt-genomes used in this study. Mt-genome sequences of 5 newly sequenced species and 24 species generated from our previous studies were highlighted in bold. (XLSX 15 kb)

Additional file 10: Optimal partitioning schemes selected by PartitionFinder for each dataset. (XLSX 10 kb)

Additional file 11: A series of in-house shell scripts used in this study. (TXT 664 bytes)

Abbreviations
BI: Bayesian inference; BIC: Bayesian information criterion; BPP: Bayesian posterior probabilities; BS: Bootstrap; ESS: Effective sample size; LBA: Long-branch attraction; MCMC: Markov Chain Monte Carlo; ML: Maximum likelihood; Mt-genome: Mitochondrial genome; NADH: Nicotinamide adenine dinucleotide; PCGs: Protein-coding genes; RY: Purine-pyrimidine

Authors' contributions
WJB and TL designed the research. HHY carried out the experiments, HHY and TL analysed the molecular data. KD identified the samples of Tingidae used in this study. The manuscript was drafted and written by HHY, WJB, TL and KD. WJB, HHY, TL and KD revised the paper. All authors have read and approved the final manuscript.

Ethics approval and consent to participate
No specific permits were required for the insects collected for this study in China. The insect specimens were collected from shrub plants and deciduous trees by net sweeping, and the field studies did not involve endangered or protected species. The species in our study are common small insects and are not included in the "List of Protected Animals in China".

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.

References
1. Cameron SL. Insect mitochondrial genomics: implications for evolution and phylogeny. Annu Rev Ent. 2014;59:95–117.
2. Bernt M, Bleidorn C, Braband A, Dambach J, Donath A, Fritzsch G, Golombek A, Hadrys H, Jühling F, Meusemann K. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. Mol Phylogenet Evol. 2013;69(2):352–64.
3. Talavera G, Vila R. What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. BMC Evol Biol. 2011;11(1):315.
4. Simon S, Hadrys H. A comparative analysis of complete mitochondrial genomes among Hexapoda. Mol Phylogenet Evol. 2013;69(2):393–403.

Yang *et al. BMC Genomics* (2018) 19:264

Page 12 of 13

5.   Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 2011;9(3):e1000602.

6.   Rota-Stabelli O, Pisani D. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst Biol. 2013;62(1):121–3.

7.   Sheffield NC, Song SL, Cameron SL, Whiting MF. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. Syst Biol. 2009;58(4):381–94.

8.   Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. Syst Entomol. 2010;35(3):429–48.

9.   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. 2010;56(2):796–807.

10. Timmermans MJTN, Barton C, Haran J, Ahrens D, Culverwell CL, Ollikainen A, Dodsworth S, Foster PG, Bocak L, Vogler AP. Family-level sampling of mitochondrial genomes in Coleoptera: compositional heterogeneity and phylogenetics. Genome Biol Evol. 2016;8(1):161–75.

11. Song F, Li H, Jiang P, Zhou X, Liu J, Sun C, Vogler AP, Cai W. Capturing the phylogeny of Holometabola with mitochondrial genome data and bayesian site-heterogeneous mixture models. Genome Biol Evol. 2016; 8(5):1411–26.

12. Li H, Shao R, Song N, Song F, Jiang P, Li Z, Cai W. Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. Sci Rep. 2015;5:8527.

13. Song N, An SH, Yin XM, Zhao T, Wang XY. Insufficient resolving power of mitogenome data in deciphering deep phylogeny of Holometabola. J Syst Evol. 2016;54(5):545–59.

14. Dowton M, Cameron SL, Austin AD, Whiting MF. Phylogenetic approaches for the analysis of mitochondrial genome sequence data in the Hymenoptera–a lineage with both rapidly and slowly evolving mitochondrial genomes. Mol Phylogenet Evol. 2009;52(2):512–9.

15. Li H, Shao R, Song F, Zhou X, Yang Q, Li Z, Cai W. Mitochondrial genomes of two Barklice, *Psococerastis albimaculata* and *Longivalvus hyalospilus* (Psocoptera: Psocomorpha): contrasting rates in mitochondrial gene rearrangement between major lineages of Psocodea. PLoS One. 2013;8(4):e61685.

16. Shao R, Kirkness EF, Barker SC. The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. Genome Res. 2009;19(5):904–12.

17. Yuan ML, Zhang QL, Zhang L, Guo ZL, Liu YJ, Shen YY, Shao R. High-level phylogeny of the Coleoptera inferred from mitochondrial genome sequences. Mol Phylogenet Evol. 2016;104:99–111.

18. Wang Y, Liu X, Winterton SL, Yan Y, Aspöck U, Aspöck H, Yang D. Mitochondrial phylogenomics illuminates the evolutionary history of Neuropterida. Cladistics. 2017;33(6):617–36.

19. Brinkmann H, Van der Giezen M, Zhou Y, De Raucourt GP, Philippe H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 2005;54(5):743–57.

20. Reyes A, Pesole G, Saccone C. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene. 2000; 259(1):177–87.

21. Simmons MP, Richardson D, Reddy AS. Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: an example from the kinesin superfamily across eukaryotes. Cladistics. 2008;24(3):372–84.

22. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452(7188):745–9.

23. Hedtke SM, Townsend TM, Hillis DM. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst Biol. 2006;55(3): 522–9.

24. Hillis DM. Inferring complex phytogenies. Nature. 1996;383(6596):130–1.

25. Li T, Hua J, Wright AM, Cui Y, Xie Q, Bu W, Hillis DM. Long-branch attraction and the phylogeny of true water bugs (Hemiptera: Nepomorpha) as estimated from mitochondrial genomes. BMC Evol Biol. 2014;14:99.

26. Delsuc F, Phillips MJ, Penny D. Comment on "hexapod origins: monophyletic or paraphyletic?". Science. 2003;301(5639):1482.

27. Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 2004;21(7):1455–8.

28. Hassanin A. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol Phylogenet Evol. 2006;38(1):100–16.

29. Breinholt JW, Kawahara AY. Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. Genome Biol Evol. 2013;5(11):2082–92.

30. Lartillot N, Philippe HA. Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004; 21(6):1095–109.

31. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009; 25(17):2286–8.

32. Husník F, Chrudimský T, Hypša V. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. 2011;9:87.

33. Morgan CC, Foster PG, Webb AE, Pisani D, Mcinerney JO, O'Connell MJ. Heterogeneous models place the root of the placental mammal phylogeny. Mol Biol Evol. 2013;30(9):2145–56.

34. Baurain D, Brinkmann H, Philippe H. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol. 2007;24(1):6–9.

35. Philippe H, Brinkmann H, Martinez P, Riutort M, Baguñà J. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS One. 2007;2(8):e717.

36. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H. Data from: additional molecular support for the new chordate phylogeny. Genesis. 2008;46(11):592–604.

37. Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 2007;7(Suppl1):S4.

38. Weirauch C, Schuh RT. Systematics and evolution of Heteroptera: 25 years of progress. Annu Rev Entomol. 2011;56:487–510.

39. Kerzhner IM. Fauna of the USSR. Bugs. Vol. 13, No. 2. Heteroptera of the family Nabidae. Leningrad: USSR Academy of Sciences, Zoological Institute, Nauka; 1981.

40. Štys P, Kerzhner IM. The rank and nomenclature of higher taxa in recent Heteroptera. Acta Entomol Bohemoslov. 1975;72(2):65–79.

41. Schuh RT, Štys P. Phylogenetic analysis of cimicomorphan family relationships (Heteroptera). J N Y Ent Soc. 1991;99:298–350.

42. Li M, Tian Y, Zhao Y, Bu W. Higher level phylogeny and the first divergence time estimation of Heteropterar (Insecta: Hemiptera) based on multiple genes. PLoS One. 2012;7(2):e32152.

43. Tian Y, Zhu W, Li M, Xie Q, Bu W. Influence of data conflict and molecular phylogeny of major clades in Cimicomorphan true bugs (Insecta: Hemiptera: Heteroptera). Mol Phylogenet Evol. 2008;47(2):581–97.

44. Wang YH, Cui Y, Rédei D, Baňař P, Xie Q, Štys P, Damgaard J, Chen PP, Yi WB, Wang Y, Dang K, Li CR, Bu WJ. Phylogenetic divergences of the true bugs (Insecta: Hemiptera: Heteroptera), with emphasis on the aquatic lineages: the last piece of the aquatic insect jigsaw originated in the late Permian/early Triassic. Cladistics. 2016;32(4):390–405.

45. Schuh RT, Weirauch C, Wheeler WC. Phylogenetic relationships within the Cimicomorpha (Hemiptera: Heteroptera): a total-evidence analysis. Syst Entomol. 2009;34(1):15–48.

46. Wang YH, Wu HY, Rédei D, Xie Q, Chen Y, Chen PP, Dong ZE, Dang K, Damgaard J, Štys P, Wu YZ, Luo JY, Sun XY, Hartung V, Kuechler SM, Liu Y, Liu HX, Bu WJ. When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera–Heteroptera. Cladistics. 2017; https://doi.org/10.1111/cla.12232.

47. Li H, Liu H, Cao L, Shi A, Yang H, Cai W. The complete mitochondrial genome of the damsel bug *Alloeorhynchus bakeri* (Hemiptera: Nabidae). Int J Biol Sci. 2012;8(1):93–107.

48. Li H, Liu H, Shi A, Štys P, Zhou X, Cai W. The complete mitochondrial genome and novel gene arrangement of the unique-headed bug *Stenopirates* sp.(Hemiptera: Enicocephalidae). PLoS One. 2012;7(1):e29419.

49. Li T, Gao C, Cui Y, Xie Q, Bu W. The complete mitochondrial genome of the stalk-eyed bug *Chauliops fallax* Scott, and the monophyly of Malcidae (Hemiptera: Heteroptera). PLoS One. 2013;8(2):e55381.

50. Yang W, Yu W, Du Y. The complete mitochondrial genome of the sycamore lace bug *Corythucha ciliata* (Hemiptera: Tingidae). Gene. 2013;532(1):27–40.

51. Kocher A, Guilbert E, Lhuillier E, Murienne J. Sequencing of the mitochondrial genome of the avocado lace bug *Pseudacysta perseae* (Heteroptera, Tingidae) using a genome skimming approach. C R Biol. 2015; 338(3):149–60.

Yang *et al. BMC Genomics* (2018) 19:264

Page 13 of 13

52. Li T, Yang J, Li Y, Cui Y, Xie Q, Bu W, Hillis DM. A mitochondrial genome of Rhyparochromidae (Hemiptera: Heteroptera) and a comparative analysis of related mitochondrial genomes. Sci Rep. 2016;6:35175.

53. Kolokotronis SO, Foox J, Rosenfeld JA, Brugler MR, Reeves D, Benoit JB, Booth W, Robison G, Steffen M, Sakas Z. The mitogenome of the bed bug *Cimex lectularius* (Hemiptera: Cimicidae). Mitochondrial DNA part B Resources. 2016;1(1):425–7.

54. Li H, Leavengood JM, Chapman EG, Burkhardt D, Song F, Jiang P, Liu J, Zhou X, Cai W. Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. Proc Royal Soc. 2017; 284:20171223.

55. Froeschner RC, Kormilev NA. Phymatidae or ambush bugs of the world: a synonymic list with keys to species, except *Lophoscutus* and *Phymata* (Hemiptera). Entomography. 1989;6:1–76.

56. Maldonado J. Systematic catalogue of the Reduviidae of the world (Insecta: Heteroptera). Caribb J Sci. Special edition. Mayagüez: University of Puerto Rico; 1990. p. 1–694.

57. Cassis G, Gross GF. Hemiptera: Heteroptera (Coleorrhyncha to Cimicomorpha). Catalogues of Australia (ed. by W. W. K. Houston and B. V. Maynard), Vol. 27.3A, p. 1–506. CSIRO Australia, Melbourne. 1995.

58. Guilbert E, Damgaard J, D'HAESE CA. Phylogeny of the lacebugs (Insecta: Heteroptera: Tingidae) using morphological and molecular data. Syst Entomol. 2014;39(3):431–41.

59. Kück P, Meid SA, Groß C, Wägele JW, Misof B. AliGROOVE – Visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support. BMC Bioinformatics. 2014;15:294.

60. Yuan ML, Zhang QL, Guo ZL, Wang J, Shen YY. The complete mitochondrial genome of *Corizus tetraspilus* (Hemiptera: Rhopalidae) and phylogenetic analysis of Pentatomomorpha. PLoS One. 2015;10(6):e0129003.

61. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 2013;62(4):611–5.

62. Bergsten J. A review of long-branch attraction. Cladistics. 2005;21(2):163–93.

63. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol Evol. 2010;2:425–40.

64. Taylor CR, Heglund NC, GMO M. Energetics and mechanics of terrestrial locomotion. I. Metabolic energy consumption as a function of speed and body size in birds and mammals. J Exp Biol. 1982;97:1–21.

65. Shen YY, Shi P, Sun YB, Zhang YP. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. Genome Res. 2009;19(10):1760–5.

66. Roques S, Fox CJ, Villasana MI, Rico C. The complete mitochondrial genome of the whiting, *Merlangius merlangus* and the haddock, *Melanogrammus aeglefinus*: a detailed genomic comparison among closely related species of the Gadidae family. Gene. 2006;383(4):12–23.

67. Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP. Adaptive evolution of energy metabolism genes and the origin of flight in bats. Proc Natl Acad Sci U S A. 2010;107(19):8666–71.

68. Tomasco IH, Lessa EP. The evolution of mitochondrial genomes in subterranean caviomorph rodents: adaptation against a background of purifying selection. Mol Phylogenet Evol. 2011;61(1):64–70.

69. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005;6(5):389–402.

70. Wallace DC. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu Rev Genet. 2005;39:359–407.

71. Schmidt T, Wu W, Goodman M, Grossman L. Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. Mol Biol Evol. 2001;18(4):563–9.

72. Zsurka G, Kudina T, Peeva V, Hallmann K, Elger CE, Khrapko K, Kunz WS. Distinct patterns of mitochondrial genome diversity in bonobos (*Pan paniscus*) and humans. BMC Evol Biol. 2010;10:270.

73. Cui Y, Xie Q, Hua J, Dang K, Zhou J, Liu X, Wang G, Yu X, Bu W. Phylogenomics of Hemiptera (Insecta: Paraneoptera) based on mitochondrial genomes. Syst Entomol. 2013;38(1):233–45.

74. Hua J, Li M, Dong P, Cui Y, Xie Q, Bu W. Comparative and phylogenomic studies on the mitochondrial genomes of Pentatomomorpha (Insecta: Hemiptera: Heteroptera). BMC Genomics. 2008;9:610.

75. Reineke A, Karlovsky P, CPW Z. Preparation and purification of DNA from insects for AFLP analysis. Insect Mol Biol. 1998;7(1):95–9.

76. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 2016;34(3):772–3.

77. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–9.

78. Foster PG. Modeling compositional heterogeneity. Syst Biol. 2004;53(3):485–95.

79. Swofford DL. PAUP v4.0b10: phylogenetic analysis using parsimony (and other methods). Sunderland: Sinauer Associates; 2002.

80. Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. J Hered. 2001;92(4):371–3.

81. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.

82. Fourment M, Gibbs MJ. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol Biol. 2006;6:1.

83. Posada D, Crandall KA. Modeltest: testing the model of DNA substitution. Bioinformatics. 1998;14(9):817–8.

84. Zhou J, Liu X, Stones DS, Xie Q, Wang G. MrBayes on a graphics processing unit. Bioinformatics. 2011;27(9):1255–61.

85. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22(21):2688–90.

86. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 2008;57(5):758–71.

87. Wu HY, Wang YH, Qiang X, Ke YL, Bu WJ. Molecular classification based on apomorphic amino acids (Arthropoda, Hexapoda): integrative taxonomy in the era of phylogenomics. Sci Rep. 2016;6:28308.

88. Marin B, Nowack ECM, Melkonian M. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 2005;156(4):425–32.