Research Article

Mitochondrial genomes infer phylogenetic relationships among the oldest extant winged insects (Palaeoptera)

Sereina Rutschmann[1,2,3], Ping Chen[4], Changfa Zhou[4], Michael T. Monaghan[1,2*]

Addresses:

[1]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany

[2]Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195 Berlin, Germany

[3]Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain

[4]The Key Laboratory of Jiangsu Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China

sereina.rutschmann@gmail.com; pingc918@hotmail.com; zhouchangfa@njnu.edu.cn; monaghan@igb-berlin.de

***Correspondence:** sereina.rutschmann@gmail.com

Department of Biochemistry, Genetics, and Immunology, University of Vigo, 36310 Vigo, Spain

Abstract

Background: The relationships among the oldest winged insects (Palaeoptera), including the Ephemeroptera (mayflies) and Odonata (dragonflies and damselflies), remain unclear. The Palaeoptera together with the Neoptera have evolved as result of a rapid divergence from a common ancestor in the distant past. Thus they are thought to be more susceptible to systematic inadequacies, including taxon sampling, choice of outgroup, marker selection, and phylogenetic methods. Here we reconstruct their phylogenetic relationship using newly sequenced mitochondrial genomes in combination with 90 additional insect mitochondrial genomes. In particular, we investigate the impact of the increased mayfly taxon sampling, the effect of rogue taxa, and the used phylogenetic framework (Bayesian inference (BI) vs. maximum likelihood (ML)) approach.

Results: We found support for the clustering of the Odonata as most ancient, extant winged insects, using BI based on an optimized data matrix. Overall, we found no support for the basal Ephemeroptera clustering and the sister relationship between the Ephemeroptera and Odonata. Our newly sequenced mitochondrial genomes of *Baetis rutilocylindratus*, *Cloeon dipterum*, and *Habrophlebiodes zijinensis* showed the complete set of 13 protein coding genes and a conserved gene orientation with the exception of two inverted tRNAs for *H. zijinensis.*

Conclusions: The increase of palaeopteran taxon sampling in combination with a Bayesian phylogenetic framework was crucial to infer phylogenetic relationships within the three ancient insect lineages of Odonata, Ephemeroptera, and Neoptera. Pruning of rogue taxa improved the number of supported nodes in all phylogenetic trees. It remains to be tested weather an increased taxon sampling might also reveal the elusive phylogenetic positions of other insect orders.

Keywords: Baetidae, Bayesian Inference, Ephemeroptera, Long-branch attraction, Mayfly, Mitochondrial genomics, Odonata, Palaeoptera problem, Polyneoptera, Pterygota, Rogue taxa

Background

Insects are the most diverse branch of metazoan life, yet the relationships among the basal branches of the winged insects (Pterygota) remains one of the major open questions in the evolution of life. The unresolved relationship of the basal orders Odonata (dragon- and damselflies) and Ephemeroptera (mayflies) to the Neoptera (i.e., to the rest of the Pterygota) has been termed the "Palaeoptera problem" [1,2]. The name Paleoptera ("old wings") reflects the inability of the Ephemeroptera and Odonata to fold their wings flat over the abdomen. This has long been considered to be the ancestral condition, in contrast to the more derived Neoptera ("new wings") (reviewed by Trautwein et al. [3]). The monophyly of the Neoptera, including the three major lineages Polyneoptera, Paraneoptera, and Holometabola, is widely accepted, although relationships among polyneopteran lineages [4] and the monophyly of the Paraneoptera [5] remain unresolved.

There are three competing hypotheses relating to the Palaeoptera problem: the Palaeoptera hypothesis ((Ephemeroptera + Odonata) + Neoptera), the basal Ephemeroptera hypothesis (Ephemeroptera + (Odonata + Neoptera)), and the basal Odonata hypothesis (Odonata + (Ephemeroptera + Neoptera)) [3,6,7]. All hypotheses are supported by morphological as well as molecular data to varying degrees (reviewed by Trautwein et al. [3]). Different authors, using the same set of genes and taxa set but different phylogenetic approaches, namely Bayesian inference (BI) vs. maximum likelihood (ML) [2,8,9], and nucleotide vs. amino acid (aa) sequences [10] obtained results supporting distinct hypotheses. The Palaeoptera hypothesis has received support from several molecular studies based on rRNA genes and nuclear DNA [5,11-13], and from nDNA phylogenomic analyses [9,14,15]. Thomas et al. [12] and Kjer et al. [11] included the most comprehensive taxon sampling with 35, respectively seven species from each of the three lineages (i.e., Ephemeroptera, Odonata, Neoptera) analyzing seven, respectively eight genes (rRNAs, nuclear DNA (nDNA), mitochondrial DNA (mtDNA)). The basal Ephemeroptera hypothesis was supported by previous studies using mitochondrial genome data [10,16,17], and a combined analysis of rRNA

69 and one nuclear gene [2]. Other studies have concluded the relationships are ambiguous [2]. Notably,
70 studies supporting the basal Ephemeroptera hypothesis have included either a limited number of
71 mayfly species (e.g., one [17,18], two [16,19], four [20], or five [10] species) or a limited number of
72 unlinked loci (e.g., one [10,16,17,20], two [18,19,21], or three [2]). The basal Odonata hypothesis
73 has received support from previous studies based on mitochondrial genome data [22,23], rRNA
74 genes [8,24-27], combining rRNA and nDNA [28], and phylogenomic nDNA [9,29]. These studies,
75 with the exceptions of few studies based on *18S* rRNA [24,25,27], included in total between two
76 [8,29] and 13 species [26] from the orders Ephemeroptera and Odonata.

77 The conflicting phylogenetic signals may result from the ancient radiation of the lineages
78 Ephemeroptera, Odonata, and Neoptera from a common ancestor in the distant past [4]. These three
79 ancient lineages appear to have diverged rapidly, leaving few characteristics to determine their
80 phylogenetic relationships [4]. Evolutionary rate heterogeneity across clades and the representation
81 of old clades by recent extant taxa make ancient relationships such as those of early winged insects
82 the most difficult challenges for phylogenetics [4,7]. Given the weak phylogenetic signal,
83 reconstructions are more susceptible to systematic errors [4,12,30-32]. Other sources of conflicting
84 signal in studies of the relationships among these three lineages may result from differences in taxon
85 sampling, sequence data, alignment methods, and phylogenetic methods used including models of
86 evolution [3,12,33]. The use of mitochondrial protein-coding genes (PCGs) derived from
87 mitochondrial genomes can circumvent some of these difficulties, since they are easier to align and
88 also appropriate models of molecular evolution are well established [34-36]. On the other hand,
89 mitochondrial genes include several drawbacks, most importantly the possible presence of
90 pseudogenes [37-39]. However, data derived from high-throughput sequencing, depending on the
91 coverage bears the potential to distinguish between functional genes and pseudogenes based on
92 higher mutation rates of the later [40].

93 Overall, mitochondrial markers are well studied and therefore the most widely employed genetic
94 markers in insects, being considered a promising "instrument" for insect systematics [41]. Insect
95 mitochondrial genomes are highly conserved, ranging from 15 to 18 kb in length, containing 37
96 genes: 13 PCGs, two ribosomal RNAs (rRNAs, *rrnL* and *rrnS*), and 22 transfer RNAs (tRNAs, *trn**)
97 [42]. A non-coding region of variable length, thought to be the origin of initiation of transcription
98 and replication, is typically present [43,44] and referred to as the AT-rich region. The typical
99 ancestral insect mitochondrial genome differs from the ancestral arthropod mitochondrial genome
100 only by the location of *trnL* [45]. Significant differences in structure, gene content, and gene
101 arrangement have been found to be the exception for highly derived taxa.

102 Importantly there is still an unbalance in the numbers of available molecular markers between the
103 different insect lineages. In particular the early winged insects (i.e., Ephemeroptera and Odonata) are
104 relatively under-represented. This is (partly) because so far most genomic data were obtained for
105 insect orders that are of economical interest (e.g., as pollinators, model species, agricultural pests,
106 and vectors of human diseases, [46]). As of January 2017, there were 39 complete or nearly complete
107 palaeopteran mitochondrial genomes (Ephemeroptera: 18, Odonata: 21) from 22 families (eleven of
108 each) available on GenBank; 22 of which were included in publications (Ephemeroptera: ten,
109 Odonata: twelve). Compared to the number of families that have been described (Ephemeroptera: 42
110 [47], Odonata: 30 [48]) this is still a relatively low number.

111 The insect order Ephemeroptera comprises over 3,000 species, comprising as the two most
112 species rich families the Baetidae (833 species) and the Leptophlebiidae (608 species) [47]. Their
113 biogeographic origin is probably Pangean, including a greater diversity and higher endemism rate in
114 the Neotropics and Australasia for the Baetidae, and the Neotropical and Afrotropical regions for the
115 Leptophlebiidae, respectively [47].

116 We investigated the relationships of the oldest extant winged insects and newly sequenced three
117 mayfly mitochondrial genomes to improve sampling of this under-represented order. New taxa

118 included one representative of the family Leptophlebiidae, for which no mitochondrial genome data
119 were available, and two representatives of the Baetidae, one from each subfamily. We used BI and
120 ML approaches and created matrices that eliminated taxa that reduced overall tree support (rogue
121 taxa) and that were affected by long-branch attraction (LBA). Our analysis included 29 palaeopteran
122 mitochondrial genomes from 20 families and 64 other insect mitochondrial genomes. We expect the
123 increased taxon sampling to improve the phylogenetic signal of the basal nodes and add support for
124 one of the three competing hypotheses.

125 **Methods**

126 **Taxon sampling**

127 We newly sequenced three mitochondrial genomes of the two most diverse Ephemeroptera families.
128 *Habrophlebiodes zijinensis* GUI, ZHANG & WU, 1996 is the first representative of the family
129 Leptophlebiidae to have its complete mitochondrial genome sequenced. *Baetis rutilocylindratus*
130 WANG, QIN, CHEN & ZHOU, 2011 and *Cloeon dipterum* L. 1761 are both members of the Baetidae
131 and representative each subfamily: *B. rutilocylindratus* from Baetinae, and *C. dipterum* from
132 Cloeninae. There was a mitochondrial genome from one other member of the Baetinae available on
133 GenBank prior to our study (*Alainites yixiani*, GQ502451, Jia & Zhou). Our analysis also included
134 four species from Archaeognatha, three from Zygentoma (formerly Thysanura), ten from Odonata,
135 16 additional species from Ephemeroptera, four from Plecoptera, four from Ensifera of Orthoptera,
136 nine from Caelifera of Orthoptera, 13 from Phasmatodea, two from Mantodea, six from Blattodea, 13
137 from Isoptera, and one each from Collembola, Dermaptera, Grylloblattodea, and Mantophasmatodea
138 (Table 1). Amino acid sequences were obtained from GenBank using a custom Python script
139 (mitogenome_ncbi.py, https://github.com/srutschmann/python_scripts).

140

141 [**Table 1** should be added here]

142

143 **Sequencing and assembly**

144 The three mayfly species were sequenced with either 454/FLX pyrosequencing (*C. dipterum*) or
145 Sanger sequencing (*B. rutilocylindratus* and *H. zijinensis*. We extracted DNA from *C. dipterum* from
146 twelve to twenty pooled reared subimago specimens, prepared a shotgun library, and sequenced four
147 lanes on a Roche (454) *GS FLX* machine at the Berlin Center for Genomics in Biodiversity Research
148 (BeGenDiv, Berlin, Germany) (see also [49]). The obtained sequence reads were trimmed and *de*
149 *novo* assembled using the software Newbler v.2.5.3 (454 Life Science Cooperation) under default
150 settings for large data sets. In order to extract the mitochondrial genome of *C. dipterum*, we
151 performed BLASTN searches [50] using as query all assembled contigs against the NCBI database.
152 We mapped all matching reads back to the mitochondrial genome with BWA [51], using the BWA-
153 SW algorithm [52] with settings suggested for 454 data by CORAL (match score = 2, mismatch
154 penalty = 2, and gap open penalty = 3, [53]).

155 Specimens of *B. rutilocylindratus* and *H. zijinensis* were collected in Zijin Hill, Nanjing, China.
156 The DNA was extracted from between two and four larvae using the DNeasy® Blood & Tissue
157 (Qiagen, Leipzig, Germany) kit. Four DNA fragments of each species were amplified with universal
158 primers (*B. rutilocylindratus*: *cob*, *cox1*, *cox3*, *rrnL*; *H. zijinensis*: *cob*, *cox1*, *nad4*, *rrnL*, [54]).
159 Subsequently, based on the previously obtained sequence information we designed six (*B.*
160 *rutilocylindratus*) respective four (*H. zijinensis*) specific primer pairs (Additional file 1: Table S1;
161 see also Li et al. [55]). Standard and long polymerase chain reactions (PCRs) were performed on a
162 DNA Engine Peltier Thermal Cycler (Bio-Rad, Shanghai, China). Therefore, we used the rTaq™
163 DNA polymerase (TaKaRa Bio, Dalian, China) for fragments smaller than two kb and the LA Taq™
164 polymerase (TaKaRa Bio, Dalian, China) for fragments larger than two kb. All PCR products were
165 purified with the Axygen agarose-out kit. When the PCR amplification signal was too weak to
166 sequence or sequencing resulted in overlap peaks, the products were ligated to pGEM®-T Easy
167 Vector (Promega, Southampton, UK) by *Escherichia coli*, and each resulting clone was sequenced.

168 All purified amplification products were sequenced successively in both directions on an ABI3130xl
169 capillary sequencer. Forward and reverse sequences were assembled and edited using CodonCode
170 Aligner v.3.5.6 (CodonCode Corporation).

171 **Sequence alignments and data matrices**

172 We prepared two data matrices of aa sequences based on the annotated mitochondrial genomes (see
173 below), including a set with all 93 taxa (all_taxa matrix), and a matrix for which we removed taxa
174 that were either identified as rogue taxa using the RogueNaRok algorithm [56,57] and showing LBA
175 [58,59] (optimized_taxa matrix; 88 taxa, see Table 1). We used GUIDANCE2 [60,61] to align the
176 sequences, and identify and remove positions detected as ambiguously aligned regions. To run
177 GUIDANCE we applied the default settings, using as multiple sequences alignment program
178 MAFFT v.7.050b (L-INS-I algorithm with default settings, [62]).

179 The best-fitting partitioning schemes and corresponding aa substitution models were estimated
180 with PartitionFinder v.2 (https://github.com/brettc/partitionfinder, [63]). We used the Bayesian
181 Information Criterion (BIC) to choose the best model, linked branch lengths, and a greedy search
182 [63] with RAxML v.8.2.8 [64]. Thereby each gene was defined as one data block (i.e., possible
183 partition).

184 To identify rogue taxa, we applied the RogueNaRok algorithm [56,57]. We used the bootstrap
185 replicates and the best supported maximum likelihood tree based on the all_taxa matrix (see below).
186 Identified rogue taxa were removed from the all_taxa matrix, resulting in the optimized_taxa matrix
187 (Table 1). Taxa showing evidence of LBA *sensu* Bergsten [65] were identified using heterogeneities
188 in sequence divergence based on a model of evolution derived from the entire data matrix. For this,
189 the Relative Composition Frequency Variability (RCFV) values of each species were calculated with
190 BaCoCa [66] using the partitioned concatenated data matrix (i.e., all_taxa matrix). Taxa with the
191 highest values were excluded in the optimized_taxa matrix, apart from the Ephemeroptera, which
192 were retained because they are the focus of our study and their monophyly is well-established [1,2].

193 **Phylogenetic reconstruction**

194 Phylogenetic reconstructions were carried out using MrBayes v.3.2.4 [67], and RAxML v.8.1.7 [64].
195 We analyzed both matrices (all_taxa, optimized_taxa) using the best-fitting partitioning scheme and
196 model for each (see Table 2). For BI, we unlinked the frequencies, gamma distributions, substitution
197 rates and the proportion of invariant sites across partitions. For each matrix, we run two independent
198 analyses of four MCMC chains with 10^7 generations, sampling every 10^3 generations and discarding
199 a burn-in of 25%. Maximum likelihood inferences were performed with 200 bootstrap replicates.
200 Trees reconstructed from the all_taxa matrix were rooted with *Tetrodontophora bielanensis*
201 (Collembola). Trees reconstructed from the optimized_taxa matrix were rooted with *Nesomachilis*
202 *australica* (Archaeognatha, according to Song et al. [10]) because *Tetrodontophora bielanensis* was
203 found to be a rogue taxon (see Results). All trees were visualized with the ape v.4.1 [68], phangorn
204 v.2.1.1 [69], and ggtree v.1.4.20 [70] packages for R v.3 [71].

205 **Annotation and characterization**

206 All mitochondrial genomes were annotated using the MITOS webserver (http://mitos.bioinf.uni-
207 leipzig.de/index.py, [72]), DOGMA [73], and MFannot (http://megasun.bch.umontreal.ca/cgi-
208 bin/mfannot/mfannotInterface.pl). The predicted PCGs were checked for stop codons and manually
209 adjusted by comparison with the longest predicted open reading frames as implemented in Geneious
210 R7 v.7.1.3 (Biomatters Ltd.), and by comparison with the respective homologous insect sequence
211 alignment. For the tRNA prediction, we additionally used ARWEN v.1.2.3 [74] and tRNAscan-SE
212 v.1.21 (http://lowelab.ucsc.edu/tRNAscan-SE/, [75]). The annotated mitochondrial genomes were
213 visualized with OrganellarGenomeDRAW (http://ogdraw.mpimp-golm.mpg.de/cgi-bin/ogdraw.pl,
214 [76]) and manually edited.

215 Nucleotide contents were retrieved using Geneious. To correct biases in the AT content due to
216 incomplete mitochondrial genomes mostly missing the AT-rich region, we removed all AT-rich
217 regions and the two rRNAs including the five close-by tRNAs (*trnL1-trnM*) and recalculated the

218 nucleotide base pair (bp) compositions. The AT and GC composition skewness were calculated as
219 follows: AT-skew = (A - T) / (A + T), and GC-skew = (G - C) / (G + C) [77].

220 **Results**

221 **Data matrices**

222 The final matrices had lengths of 3,675 aa (all_taxa matrix, including 93 taxa) and 3,673 aa
223 (optimized_taxa matrix, including 88 taxa, see Table 1). As best-fitting partitioning scheme for the
224 two matrices, we identified either five partitions (all_taxa matrix, Table 2) or four partitions
225 (optimized_taxa, Table 2). As best-fitting models of aa sequence evolution, we identified MtZoa [78]
226 and MtArt [35].

227

228 **Table 2** Overview of data matrices, including the best-fitting partitioning and model schemes, and
229 amino acid (aa) sequence length

Matrix	Model	Length [aa]
all_taxa		
partition_1: *atp6*, *atp8*, *nad2*, *nad3*, *nad6*	MTZOA + Γ + I + F	840
partition_2: *cox1*	MTZOA + Γ + I	528
partition_3: *cob*, *cox2*	MTZOA + Γ + I + F	608
partition_4: *cox3*	MTART + Γ + I	268
partition_5: *nad1*, *nad4*, *nad4L*, *nad5*	MTZOA + Γ + I + F	1431
optimized_taxa		
partition_1: *atp6*, *atp8*, *nad2*, *nad3*, *nad6*	MTZOA+ Γ + I + F	843
partition_2: *cox1*	MTZOA + Γ + I	529
partition_3: *cob*, *cox2*, *cox3*	MTZOA + Γ + I + F	878
partition_4: *nad1*, *nad4*, *nad4L*, *nad5*	MTZOA + Γ + I + F	1423

230

231 The RogueNaRok analysis identified the outgroup *T. bielanensis* as well as *Gryllotalpa orientalis*
232 (Orthoptera) and *Phraortes* sp. (Phasmatodea) as taxa with uncertain phylogenetic positions, leading
233 to less accurate phylogenetic reconstructions. *G. orientalis* (Orthoptera) and *Phraortes* sp.

234 (Phasmatodea) clustered within the corresponding order using the all_taxa matrix, although the
235 nodes containing these species were not supported by ML. In the test for LBA, *Aposthonia japonica*
236 (Embioptera) and *Challia fletcheri* (Dermaptera) had high RCFV values (0.0237 and 0.0307,
237 respectively) suggesting they were affected by LBA. Both species clustered within other orders:
238 *Challia fletcheri* (Dermaptera) within the Ephemeroptera as sister taxon to a clade containing all
239 representatives of the family Baetidae, and *A. japonica* (Embioptera) within the Phasmotodea as
240 sister taxon to a clade containing all species of the order with the exception of *T. californicum* (Table
241 3, Additional file 2: Figure S1). The mayflies Teloganodidae sp. (0.0305) and the three baetids *A.*
242 *yixiani* (0.0235), *B. rutilocylindratus* (0.0232), and *C. dipterum* (0.0201) showed relatively high
243 RCFV values.

244 **Phylogenetic reconstruction**

245 The Bayesian reconstruction based on the optimized_taxa matrix produced the tree with most
246 support compared to ML or to either analysis of the all_taxa matrix (Figure 1, Table 3, Additional
247 files 2-4: Figures S1-S3). The basal Odonata hypothesis was highly supported in this tree (Bayesian
248 posterior probability (BPP) = 1.00), whereas neither the basal Ephemeroptera hypothesis nor the
249 Palaeoptera hypothesis received support in any analysis (Table 3). The Odonata were monophyletic
250 in all analyses (BPP = 1.00, BS = 100%) whereas the Ephemeroptera were monophyletic only using
251 the optimized_taxa matrix (Figure 1, Table 3, Additional file 3: Figures S2).

252

253 **Table 3** Support for the three hypotheses and node support for clades of interest. Node support is
254 based on Bayesian inference (BI) and maximum likelihood (ML) analyses of two data matrices:
255 optimized_taxa and all_taxa, whereby optimized_taxa had 5 terminal nodes removed based on the
256 occurrence of rogue taxa and long branch attraction. Asterices indicate hypothesis/node support (**
257 = Bayesian posterior probability (BPP) ≥ 0.95 and Bootstrap support (BS) = 90%) * = BPP ≥ 0.90

258 and BS ≥ 80%). Superorders are underlined, and suborders are italicised. Letters in brackets refer to

259 monophyletic clades in the Figure 1

Hypothesis/Clade	BI		ML	
	optimized_taxa	all_taxa	optimized_taxa	all_taxa
Palaeoptera hypothesis				
Ephemeroptera hypothesis				
Odonata hypothesis	**			
Winged insects (A)	**			
<u>Palaeoptera</u>				
Odonata (B)	**	**	**	**
Ephemeroptera (C)	**		**	
<u>Polyneoptera</u>			*	
Plecoptera (D)	**	**	**	**
Orthoptera				
Ensifera (E)	**	**	**	
Caelifera (F)	**	**	**	**
Phasmatodea[a] (G)	**	**	**	**
<u>Dictyoptera (H)</u>	**	**	**	**
Mantodea (I)	**	**	**	**
Isoptera (J)	**	**	**	**

260 [a], Includes *Aposthonia japonica* (Embioptera) in reconstructions using the all_taxa matrix.

261

262 Most other nodes of interest were consistently supported across different matrices and analyses

263 (Table 3). The Polyneoptera were moderately supported (BS ≥ 80%) for the optimized_taxa matrix

264 using the ML. Within the Polyneoptera, all orders except Orthoptera and Blattodea were

265 monophyletic (BPP ≥ 0.95, and BS ≥ 90%, Table 3). The two Orthoptera suborders Ensifera and

266 Caelifera were monophyletic (Table 3), whereby the Ensifera clustered together with the Plecoptera

267 for the optimized_taxa matrix using both ML and BI. The Grylloblattodea, Mantophasmatodea, and

268 Phasmatodea were all well supported using the optimized_taxa matrix. For the all_taxa matrix also

269 the Embioptera were recovered in this clade as sister taxa to the Phasmatodea excluding *Timema*

270 *californicum* (Figure 1, Additional files 2, 4: Figures S1, S3). The superorder Dictyoptera was
271 monophyletic clade, consisting of the monophyletic Mantodea and the Blattodea including the
272 Isoptera.

273 **Mitochondrial genomes**

274 The mitochondrial genome sequences have been deposited at GenBank with the accession numbers
275 XXXXXXXX, XXXXXXXX, and XXXXXXXX. The pyrosequencing run of *C. dipterum* using the
276 454 GS FLX system resulted in 651,306 reads, of which 1.14% mapped to the assembled
277 mitochondrial genome. The depth of coverage for the *C. dipterum* mitochondrial genome was 249.2x
278 (± 80.6 SD, Figure 2a). The three mayfly mitochondrial genomes were 15,407 bp (*C. dipterum*),
279 14,883 bp (*B. rutilocylindratus)*, and 14,355 bp (*H. zijinensis*) long (Table 4). For *B.*
280 *rutilocylindratus*, the complete mitochondrial genome was sequenced (Figure 2b). The AT-rich
281 regions were incomplete for genomes of *C. dipterum* and *H. zijinensis*. The three tRNAs between the
282 AT-rich region and *nad2* were also missing in *H. zijinensis* due to the incomplete sequencing (Figure
283 2c). All three sequenced mitochondrial genomes contained the entire set of 13 PCGs, two rRNAs,
284 and either 22 tRNAs (*B. rutilocylindratus*, *C. dipterum*) or 19 for the incomplete mitochondrial
285 genome of *H. zijinensis*, including 23 coded at the (+) strand (21 for *H. zijinensis*) and 14 at the (-)
286 strand (13 for *H. zijinensis*) (Figure 2).

287 The order of the genes was conserved, with the exceptions of the two inverted tRNAs Arginine
288 and Alanine for *H. zijinensis* (Figure 2c). All PCGs started with the ATN codons (ATT, ATG, ATA),
289 and mostly ended with the complete termination codon (TAA or TAG). Exceptions were an
290 incomplete T termination codon for *nad4* in *B. rutilocylindratus* and *cox1* started with CTC in *C.*
291 *dipterum*. The AT content of the mitochondrial genomes that were corrected for the missing regions,
292 was lowest in *B. rutilocylindratus* (59.9%, Table 4, Additional file 5: Table S2). The mean AT
293 content for all available mayfly mitochondrial genomes was 65.34%. The average AT-skew was -
294 0.03 (± 0.04 SD) and the average GC-skew was -0.13 (± 0.14 SD).

295

296 **Table 4** Overview of sequenced ephemeropteran mitochondrial genomes. Total sequence length in
297 base pairs (bp) and individual nucleotide compositions of mayfly mitochondrial genomes calculated
298 based on the whole available sequences and the ones corrected for the incomplete mitochondrial
299 genomes (see Methods; for the complete list of all ephemeropteran mitochondrial genomes see
300 Additional file 5: Table S2)

Family	Species	Length [bp]	A%	C%	G%	T%	GC%	AT%
Baetidae	*Baetis rutilocylindratus*	14,883[a]	27.3	19.8	20.1	32.8	39.9	60.1
		12,234	26.7	19.8	20.4	33.2	40.2	59.9
Baetidae	*Cloeon dipterum*	14,355	30.6	14.7	16.3	38.4	30.9	69.0
		12,197	29.9	14.5	16.9	38.7	31.4	68.6
Leptophlebiidae	*Habrophlebiodes zijinensis*	15,407	33.8	20.2	11.0	35.1	31.2	68.9
		12,417	33.6	20.6	11.5	34.3	32.1	67.9

301 [a], Indicates complete mitochondrial genomes; rest are incomplete mitochondrial genomes.

302

303 **Discussion**

304 **Phylogenetic relationships of Palaeoptera**

305 Our Bayesian reconstruction based on the optimized_taxa matrix provided strong support for the
306 basal Odonata hypothesis. Previous work based on mitochondrial genome data is (partially)
307 congruent with our findings [10,22,23]. In contrast to our results, a handful of studies based on
308 nucleotide and aa sequences of mitochondrial PCGs and mitochondrial RNA (mtRNAs) found
309 overall strong support for the basal Ephemeroptera hypothesis [10,16,17,20]. Interestingly, Song et
310 al. [10], supporting overall the Ephemeroptera hypothesis, found evidence for the Odonata
311 hypothesis when using aa sequence data. Our results confirm the trend towards a better support for
312 the ancestral position of the Odonata when using an increased mayfly taxon sampling (more than ten
313 species per lineage) [24,25,27]. The exceptions for this are the work by Ogden and Whiting [2] and

314 Thomas et al. [12], that overall support the Ephemeroptera respective Palaeoptera hypothesis but
315 contain some data sets in favor of the Odonata hypothesis. Early inconsistencies based on different
316 sequence alignment strategies [2] appear to become less relevant due to the advances of multiple
317 sequence alignment programs. Overall, the vast majority of studies were based on BI and ML.
318 Thomas et al. [12] recovered increased resolution using BI in comparison to ML. Different data
319 partitioning (i.e., gene trees resulting in different tree topologies) by Kjer et al. [11] supported
320 different hypotheses for different genes. Evidently, using rRNAs (nDNA and mtDNA) and the
321 nDNA EF1-α the Odonata hypothesis was recovered whereas the analyses based on the combined
322 and on the mtDNA data matrix resulted in a monophyletic palaeopteran clade. It is questionable
323 whether an increase in analyzed genomic data (i.e., phylogenomic studies) will be able to resolve the
324 Palaeoptera problem. Notably, Misof et al. [79] and Regier et al. [15], using 1,478 PCGs and seven
325 palaeopteran species, respective 62 PCGs and four palaeoptera species, recovered the Ephemeroptera
326 and Odonata as sister clades (i.e., the Palaeoptera hypothesis) but without high support values.
327 Overall, more knowledge about the use and limitation of individual markers are needed. For example
328 Simon et al. [80] found that proteins involved in cellular processes and signaling harbor the most
329 phylogenetic signal.

330 The increased mayfly taxon sampling highlighted the phylogenetic diversity (as indicated by the
331 branch lengths of the tip taxa, Figure 1, Additional files 2-4: Figures S1-S3) of the Palaeoptera. The
332 overall long-branch lengths and high rate-heterogeneity across sites might be one of the explanations
333 for the "Palaeoptera problem". Especially the family of the Baetidae, which was long thought to be
334 the most ancestral mayfly family [81], showed high base composition heterogeneity possibly
335 evidencing LBA. However, in agreement with other more recent studies we also recovered *S.*
336 *chinensis* as sister taxa to all other mayflies [55,82]. It remains to be resolved whether an increased
337 taxon sampling may obviate LBA *sensu* Bergsten [65].

338 **Phylogenetic relationships of Polyneoptera**

339 Besides the Palaeoptera, a universal consensus on the placement of the Plecoptera, Dermaptera,
340 Embioptera, and Xenoptera (i.e., Grylloblattodea and Mantophasmatodea) remains elusive [10]. The
341 monophyly of the Polyneoptera has become more widely accepted [5,10,23,79,83,84], although
342 several studies and data matrices do not support their monophyly [10,11,80]. This is also reflected by
343 our results, recovering the Polyneoptera as moderately supported monophyletic clade when using the
344 optimized_taxa matrix with ML. The placement of the order Plecoptera also remains elusive. The
345 order often clusters among the more ancestral polyneopteran orders [10,23,79] and in a close
346 relationship to the Dermaptera, although earlier studies using few species and the aa sequences of the
347 mitochondrial PCGs also recovered a sister relationship to the Ephemeroptera [22,85]. Notably, Song
348 et al. [10] was the only study including four dermapteran species. The monophyly of the Orthoptera
349 has been established by previous studies based on mitochondrial genome data and a large number of
350 PCGs [10,79,86,87]. However, in a recent study, the Ensifera also clustered as more ancestral clade
351 [10]. As for this study, Song et al. [10] also used BI and ML based on mitochondrial genome data.
352 Notably, Song et al. [10] explain the more ancestral position of Ensifera by more similar sequence
353 compositional bias in comparison to the outgroups. The relationship between Grylloblattodea and
354 Mantophasmatodea (i.e., Xenoptera superorder) remains elusive. Mostly this is due to the limited
355 amount of available molecular data, including ordinal-level phylogenies based on two [5,10] or three
356 xenopteran species [79]. Using mitochondrial genome data and one species per order, also Song et al.
357 [10] did not recover them as a monophyletic clade. Their monophyly was supported by previous
358 studies using one Mantophasmatodea and one Grylloblattodea species [5], respective one
359 Mantophasmatodea and two Grylloblattodea species [79]). The Phasmatodea have also been found in
360 previous studies as paraphyletic clade due to the outside position of *T. californicum* [10,16]. In
361 contrast, Misof et al. [79], using a large set of nuclear PCGs also found *Timema* as the sister species
362 to all other representatives of the order. The monophyly of the superorder Dictyoptera is generally

363 accepted and the hierarchical clustering (Mantodea + (Blattodea + Isoptera)) has been well supported
364 [5,10,11,16,23,79,88,89] using nucleotide and aa sequences of both mtDNA and nDNA.

365 **Optimizing data matrix**

366 The optimization of the data matrix overall resulted in a better-supported tree (as measured by the
367 total number of supported nodes and the fact that all nodes supported in the all_taxa matrix were also
368 recovered in the optimized_taxa matrix). In a study on butterfly phylogenetics, the authors also
369 removed rogue taxa and found dramatically increased bootstrap support values [90]. Comparing the
370 number of supported nodes from the BI and ML approach, the former always resulted in more
371 resolved nodes. This result is consistent with previous work on the "Palaeoptera problem",
372 recovering more support for analyses using BI [12]. Generally, the choice of the outgroup species
373 was reported as being crucial for resolving problematic splits in the tree of life such as insects origin
374 [12]. Thus, finding the outgroup as a rogue taxon was perhaps not surprising. The close phylogenetic
375 relationship of the Dermaptera and Ephemeroptera (Additional files 2, 4: Figures S1, S3) might be
376 misleading due to LBA. Li et al. [55] also found the order Dermaptera as being closely related to the
377 mayflies. On the other hand studies based on phylogenomic data reported the Dermaptera as sister
378 taxa to the orders Plecoptera [5,16,23,80,91] Zoraptera [28,79]. Using mitochondrial genome data
379 and an increased taxon sampling (i.e., four dermapteran species), Song et al. [10] found the
380 Dermaptera and/or Plecoptera being sister to the remaining taxa within Polyneoptera. Notably, when
381 only using one species of Dermaptera, *C. fletcheri* emerged as sister taxa to the Ephemeroptera [10].
382 This result in agreement with our findings is presumably due to LBA. Phylogenomic data matrices
383 (e.g., expressed sequence tags data: [83], 1,478 PCGs: [79]) and nuclear genes [5] found the
384 Embioptera to be sister clade to the Phasmatodea. Using mitochondrial genomes data and BI, the
385 Embioptera were also found as sister group to the Phasmatodea (excluding *T. californicum*) [10].
386 However, in the same study, they also found support for the sister-relationship of Embioptera-
387 Zoraptera using ML [10]. Overall, more phylogenetic studies will be needed to clarify this issue;

388 mostly also because the use of few taxa from anomalous orders (i.e., Dermaptera and Embioptera)
389 tend to evoke LBA [10].

390 **Characterization of ephemeropteran mitochondrial genomes**

391 Mayfly mitochondrial genomes are widely conserved in gene content and nucleotide composition
392 across eleven families, with only a few differences in the content of their tRNAs. The coverage of
393 the *Cloeon*-mitochondrial genome was in agreement with other studies, using the same sequencing
394 platform (e.g., 59-281x, [92]). The missing of part of the AT-rich region is due to reduced
395 sequencing and assembly efficiency of this low complexity region and common among insect
396 mitochondrial genomes [55,93]. The gene order and orientation were with the exception of two
397 shifted tRNAs for *H. zijinensis* identical to the ancestral insect mitochondrial genome [41,42,54].
398 Other mayflies are also reported to miss complete T termination codons in the genes *cox2* and *nad5*
399 [55,93,94]. The two rRNAs were located between *trnL* and *trnV* (*rrnL*), and between *trnV* and the
400 AT-rich region (*rrnS*), respectively. All ephemeropteran mitochondrial genomes contain an AT-rich
401 region, which is placed between the *rrnS* (- strand) and *trnI* (+ strand). Li et al. [55] reported two
402 distinct parts within the AT-rich region in *Siphluriscus chinensis,* which also seems to be present in
403 *C. dipterum*. Therein, they described the so called CR_1, which is located close to the *rrnS* and has a
404 high AT content (71.6%), including six identical 140 bp sequences, and the CR_2, which is close to
405 the *trnI* and has a lower AT content (58.1%). *E. orientalis* contains two identical 55 bp long
406 sequences in the AT-rich region [94]. Few mayfly mitochondrial genomes differ in their gene
407 content from the ancestral insect mitochondrial genome, possessing one additional tRNA. The two
408 heptageniid species *Parafronurus youi* and *Epeorus* sp. encode a second copy of the *trnM* (AUG,
409 *trnM2*) gene located between *trnI* and *trnQ* [17,93]. For *S. chinensis*, an additional *trnK2* (AAA)
410 gene is described [55]. Overall, Song et al. [10] found similar genetic distances for the early
411 branching insects lineages. The AT contents for the Ephemeroptera were very similar to the ones of
412 Odonata (62.6%-68.5% for *cox1*, [95]).

Conclusions

Our analysis based on an increased mayfly taxon sampling and data matrix optimization supported the Odonata hypothesis and unravelled fine structural changes within the palaeopteran mitochondrial genomes. While both BI and ML overall resulted in highly supported trees, only the BI based on an optimized taxa matrix highly supported the sister-relationship of the Ephemeroptera and Neoptera. The optimized taxa matrix, excluding rogue taxa and taxa with LBA, resulted in an overall better supported tree (as measured by the number of supported nodes) for both BI and ML. Our findings highlight the essential need to increase the taxon sampling of under-represented lineages (such as the Palaeoptera) in order to resolve their phylogenetic position. Here we demonstrated the need for an increased taxon sampling in combination with data matrix optimization, and the use of different phylogenetic approaches in order to resolve an ancient radiation. Establishing general recommendation for data matrix optimization requires additional analyses on a broader range of lineages.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Newly generated sequences are available on GenBank (XXXXXXXX, XXXXXXXX, and XXXXXXXX).

Competing interests

The authors declare that they have no competing interests.

Funding

Research was funded by the Leibniz Association PAKT für Forschung und Innovation ('FREDIE' project: SAW-2011-ZFMK-3).

Authors' contributions

SR and MTM conceived the study; SR sequenced and analyzed the *Cloeon dipterum* genome, and performed all combined analyses; PC and CZ obtained and analyzed the *Baetis rutilocylindratus* and *Habrophlebiodes zijinensis* genomes; SR and MTM wrote the manuscript. All authors made contributions to subsequent revisions and agreed to the final version.

Acknowledgements

We are grateful to D. H. Funk from the Stround Water Research Center for providing the *Cloeon dipterum* specimens, S. Mbedi, K. Preuß, and L. Wächter for their great help with laboratory work, G. Glöckner, and C. Mazzoni for constructive comments about the analysis of the mitochondrial genomes, and Zedat-HPC at the Freie Universität Berlin, Germany for providing access to high-performance computing clusters. SR thanks the Janggen-Pöhn-Stiftung (http://www.janggen-poehn.ch/) for a Postdoctoral stipend. We also want to thank the members of our research groups for constructive working environments.

457 **Tables**

458 **Table 1** Set of mitochondrial genomes with according GenBank accession numbers

Order	Family	Species	Accession
Archaeognatha	Machilidae	*Pedetontus silvestrii*	NC_011717
Archaeognatha	Machilidae	*Petrobius brevistylis*	NC_007688
Archaeognatha	Machilidae	*Songmachilis xinxiangensis*	NC_021384
Archaeognatha	Machilidae	*Trigoniophthalmus alternatus*	NC_010532
Archaeognatha	Meinertellidae	*Nesomachilis australica*	NC_006895
Blattodea	Blattidae	*Periplaneta americana*	NC_016956
Blattodea	Blattidae	*Periplaneta fuliginosa*	NC_006076
Blattodea	Corydiidae	*Eupolyphaga sinensis*	NC_014274
Blattodea	Cryptocercidae	*Cryptocercus relictus*	NC_018132
Blattodea	Ectobiidae	*Blattella bisignata*	NC_018549
Blattodea	Ectobiidae	*Blattella germanica*	NC_012901
Collembola	Tetrodontophorinae	*Tetrodontophora bielanensis*[a]	NC_002735
Dermaptera	Pygidicranidae	*Challia fletcheri*[a]	NC_018538
Embioptera	Oligotomidae	*Aposthonia japonica*[a]	AB639034
Ephemeroptera	Ameletidae	*Ameletus* sp1	KM244682
Ephemeroptera	Baetidae	*Alainites yixiani*	NC_020034
Ephemeroptera	Baetidae	*Baetis rutilocylindratus*	XXXXXXXX
Ephemeroptera	Baetidae	*Cloeon dipterum*	XXXXXXXX
Ephemeroptera	Caenidae	*Caenis pycnacantha*	GQ502451
Ephemeroptera	Ephemerellidae	*Ephemerella* sp.	KM244691
Ephemeroptera	Ephemerellidae	*Vietnamella dabieshanensis*	NC_020036
Ephemeroptera	Ephemerellidae	*Vietnamella* sp.	KM244655
Ephemeroptera	Ephemeridae	*Ephemera orientalis*	NC_012645
Ephemeroptera	Heptageniidae	*Epeorus* sp.	KM244708
Ephemeroptera	Heptageniidae	*Paegniodes cupulatus*	NC_020035
Ephemeroptera	Heptageniidae	*Parafronurus youi*	NC_011359
Ephemeroptera	Isonychiidae	*Isonychia ignota*	NC_020037
Ephemeroptera	Leptophlebiidae	*Habrophlebiodes zijinensis*	XXXXXXXX
Ephemeroptera	Potamanthidae	*Potamanthus* sp.	KM244674
Ephemeroptera	Siphlonuridae	*Siphlonurus immanis*	NC_013822
Ephemeroptera	Siphlonuridae	*Siphlonurus* sp.	KM244684
Ephemeroptera	Siphluriscidae	*Siphluriscus chinensis*	HQ875717
Ephemeroptera	Teloganodidae	sp.	KM244670; KM244703[b]
Grylloblattodea	Grylloblattidae	*Grylloblatta sculleni*	DQ241796
Isoptera	Hodotermitidae	*Microhodotermes viator*	NC_018122
Isoptera	Kalotermitidae	*Neotermes insularis*	NC_018124
Isoptera	Rhinotermitidae	*Coptotermes lacteus*	NC_018125
Isoptera	Rhinotermitidae	*Heterotermes* sp.	NC_018127

Isoptera	Rhinotermitidae	*Reticulitermes flavipes*	NC_009498
Isoptera	Rhinotermitidae	*Reticulitermes hageni*	NC_009501
Isoptera	Rhinotermitidae	*Schedorhinotermes breinli*	NC_018126
Isoptera	Termitidae	*Drepanotermes* sp.	NC_018129
Isoptera	Termitidae	*Macrognathotermes errator*	NC_018130
Isoptera	Termitidae	*Macrotermes subhyalinus*	NC_018128
Isoptera	Termitidae	*Nasutitermes triodiae*	NC_018131
Isoptera	Termopsidae	*Porotermes adamsoni*	NC_018121
Isoptera	Termopsidae	*Zootermopsis nevadensis*	NC_024658
Mantodea	Caliridinae	*Leptomantella albella*	NC_024028
Mantodea	Mantidae	*Tamolanica tamolana*	NC_007702
Mantophasmatodea	Mantophasmatidae	*Sclerophasma paresisensis*	NC_007701
Odonata	Calopterygidae	*Vestalis melania*	NC_023233
Odonata	Coenagrionidae	*Ischnura pumilio*	NC_021617
Odonata	Corduliidae	*Cordulia aenea*	JX963627
Odonata	Epiophlebiidae	*Epiophlebia superstes*	NC_023232
Odonata	Euphaeidae	*Euphaea formosa*	NC_014493
Odonata	Gomphidae	*Davidius lunatus*	NC_012644
Odonata	Gomphidae	*Ictinogomphus* sp.	KM244673
Odonata	Libellulidae	*Hydrobasileus croceus*	KM244659
Odonata	Libellulidae	*Orthetrum triangulare melania*	AB126005
Odonata	Pseudolestidae	*Pseudolestes mirabilis*	NC_020636
Orthoptera	Acrididae	*Acrida willemsei*	NC_011303
Orthoptera	Acrididae	*Calliptamus italicus*	NC_011305
Orthoptera	Acrididae	*Gomphocerus sibiricus*	NC_021103
Orthoptera	Acrididae	*Oedaleus decorus asiaticus*	NC_011115
Orthoptera	Gryllotalpidae	*Gryllotalpa orientalis*[a]	NC_006678
Orthoptera	Pneumoridae	*Tanaocerus koebelei*	NC_020777
Orthoptera	Prophalangopsidae	*Tarragoilus diuturnus*	NC_021397
Orthoptera	Pyrgomorphidae	*Atractomorpha sinensis*	NC_011824
Orthoptera	Rhaphidophoridae	*Troglophilus neglectus*	NC_011306
Orthoptera	Romaleidae	*Xyleus modestus*	NC_014490
Orthoptera	Tetrigidae	*Tetrix japonica*	NC_018543
Orthoptera	Tettigoniidae	*Gampsocleis gratiosa*	NC_011200
Orthoptera	Tettigoniidae	*Ruspolia dubia*	NC_009876
Phasmatodea	Bacillidae	*Bacillus rossius*	GU001956
Phasmatodea	Diapheromeridae	*Micadina phluctainoides*	NC_014673
Phasmatodea	Diapheromeridae	*Sipyloidea sipylus*	AB477470
Phasmatodea	Heteropterygidae	*Heteropteryx dilatata*	NC_014680
Phasmatodea	Heteropterygidae	*Orestes mouhotii*	AB477462
Phasmatodea	Phasmatidae	*Entoria okinawaensis*	NC_014694
Phasmatodea	Phasmatidae	*Extatosoma tiaratum*	NC_017748
Phasmatodea	Phasmatidae	*Megacrania alpheus adan*	NC_014688

Phasmatodea	Phasmatidae	*Neohirasea japonica*	AB477469
Phasmatodea	Phasmatidae	*Phobaeticus serratipes*	NC_014678
Phasmatodea	Phasmatidae	*Phraortes* sp.[a]	NC_014705
Phasmatodea	Phasmatidae	*Ramulus irregulariterdentatus*	NC_014702
Phasmatodea	Timematidae	*Timema californicum*	DQ241799
Plecoptera	Perlidae	*Dinocras cephalotes*	NC_022843
Plecoptera	Perlidae	*Kamimuria wangi*	NC_024033
Plecoptera	Perlidae	*Togoperla* sp.	KM409708
Plecoptera	Pteronarcyidae	*Pteronarcys princeps*	NC_006133
Zygentoma	Lepidotrichidae	*Tricholepidion gertschi*	NC_005437
Zygentoma	Lepismatidae	*Thermobia domestica*	NC_006080
Zygentoma	Nicoletiidae	*Atelura formicaria*	NC_011197

459 [a], Taxa were excluded in the optimized_taxa matrix (see Methods).
460 [b], Represents sequence consisting of two genomic fragments (contigs) from the same sample.

461

462

463

464

465

466

467

468

469

470

Figures

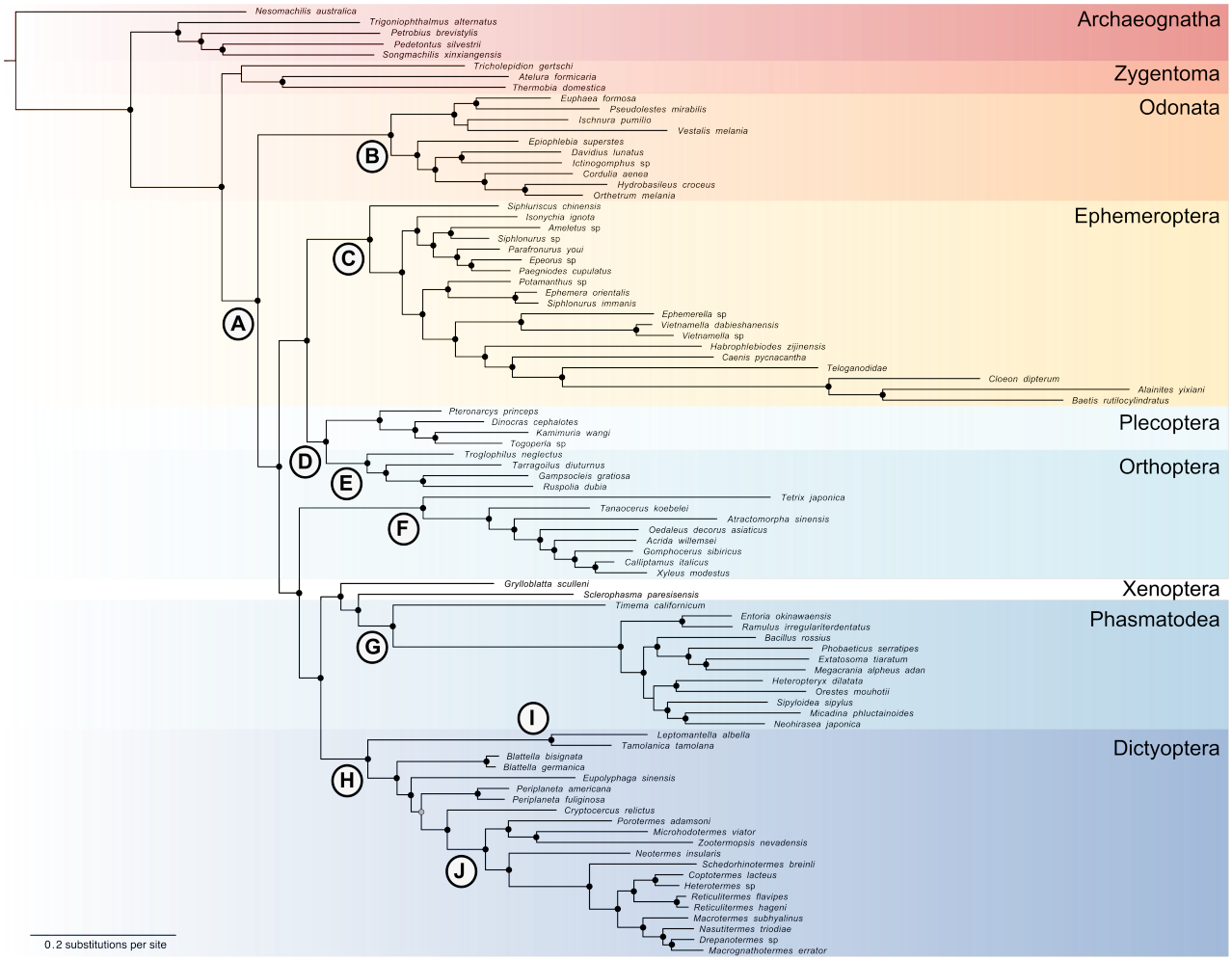
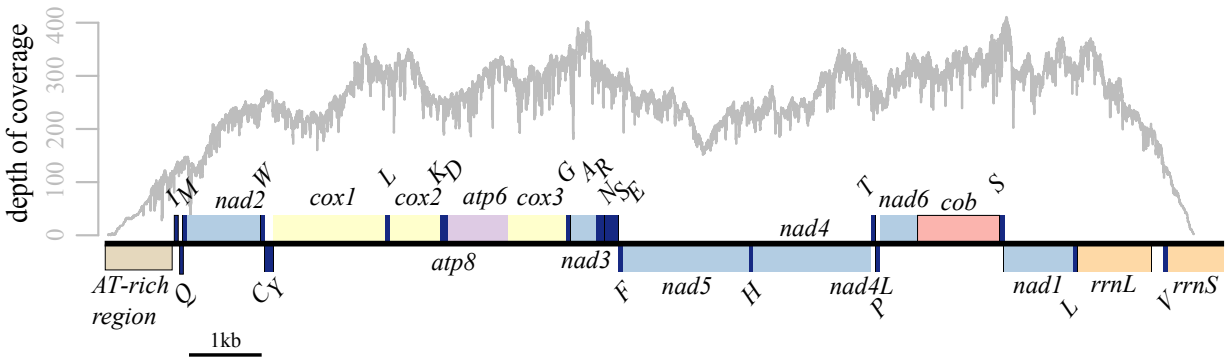
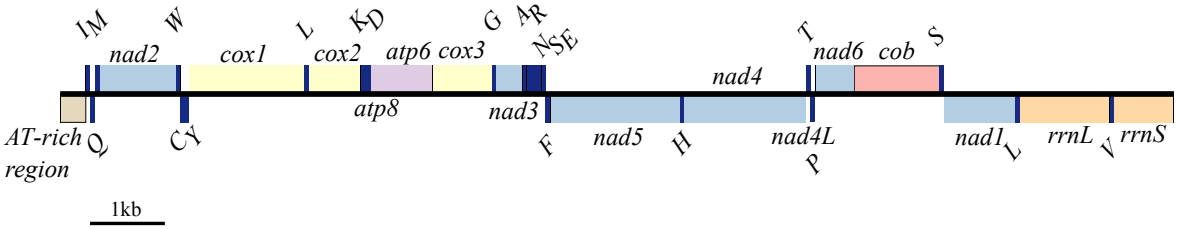


Fig. 1 Phylogenetic relationships of insect major orders using Bayesian inference reconstruction based on mitochondrial genome data using the concatenated amino acid sequences of the optimized_taxa matrix (see Table 1). Filled circles indicate supported nodes; whereby black circles represent Bayesian posterior probability (BPP) = 1, and grey circles BPP ≥ 0.95. Letters indicate clades mentioned in the text (see also Table 3)

(a) ***Cloeon dipterum*** L. 1761



(b) ***Baetis rutilocylindratus*** Wang, Qin, Cheng & Zhou, 2011



(c) ***Habrophlebiodes zijinensis*** Gui, Zhang & Wu, 1996

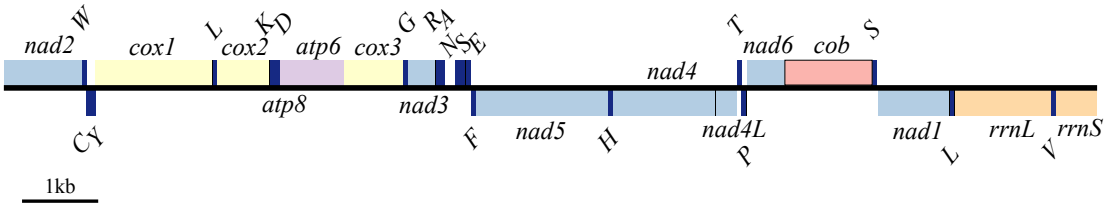


Fig. 2 Mitochondrial genome maps of three sequenced Ephemeroptera species. The nearly complete mitochondrial genome of (**a**) *Cloeon dipterum*, including coverage depth, the complete mitochondrial genome of (**b**) *Baetis rutilocylindratus,* and the incomplete mitochondrial genome of (**c**) *Habrophlebiodes zijinensis*. Transfer RNA genes are indicated by single-letter IUPAC-IUB abbreviations for their corresponding amino acid. Protein coding genes and ribosomal RNA genes are listed and colored in the following way: *atp6*, *atp8*, ATP synthase subunits 6 and 8 genes (violet); *cob*, cytochrome oxidase *b* gene (red); *cox1-cox3*, cytochrome oxidase *c* subunit 1-3 genes (yellow); *nad1-6, nad4L*, NADH dehydrogenase subunits 1-6 and 4L (light blue); *rrnS, rrnL*, small and large ribosomal RNA subunits (orange); AT-rich region (brown). Genes located at the (+) strand appear above the central line. Genes located on the (-) strand appear below the central line.

490 **Additional files**

491 **Additional file 1: Table S1.** Universal primers for Sanger sequencing. List of primer pairs used for
492 the PCR amplification and Sanger sequencing of *Baetis rutilocylindratus* (Baetidae) and
493 *Habrophlebiodes zijinensis* (Leptophlebiidae).

494 **Additional file 2: Fig. S1.** Phylogenetic relationships of insect major orders using Bayesian
495 inference reconstruction based on mitochondrial genome data using the concatenated amino acid
496 sequences of the all_taxa matrix (see Table 1). Filled circles indicate well-supported nodes; whereby
497 black circles represent Bayesian posterior probability (BPP) = 1, and grey circles BPP ≥ 0.95. Scale
498 bar indicates substitutions per site. Coloured tip labels refer to the orders, whereby light blue =
499 Archaeognatha (AR), dark blue = Blattodea (BL), light green = Ephemeroptera (EP), dark green =
500 Grylloblattodea (GR), red = Isoptera (IS), light orange = Mantodea (MT), dark red =
501 Mantophasmatodea (MP), dark orange = Odonata (OD), light purple = Orthoptera (OR), dark purple
502 = Phasmatodea, yellow = Plecoptera (PL), and brown = Zygentoma (ZY). Five taxa being excluded
503 from the optimized_taxa matrix are coloured in dark grey.

504 **Additional file 3: Fig. S2.** Phylogenetic relationships of insect major orders using maximum
505 likelihood reconstruction based on mitochondrial genome data using the concatenated amino acid
506 sequences of the optimized_taxa matrix (see Table 1). Filled circles indicate well-supported nodes;
507 whereby black circles represent Bootstrap Support (BS) ≥ 90, and grey circles BS ≥ 80. Scale bar
508 indicates substitutions per site. Coloured tip labels refer to the orders (for details see Additional file
509 2: Figure S1).

510 **Additional file 4: Fig. S3.** Phylogenetic relationships of insect major orders using maximum
511 likelihood reconstruction based on mitochondrial genome data using the concatenated amino acid
512 sequences of the all_taxa matrix (see Table 1). Filled circles indicate well-supported nodes; whereby
513 black circles represent Bootstrap Support (BS) ≥ 90, and grey circles BS ≥ 80. Scale bar indicates

514 substitutions per site. Coloured tip labels refer to the orders with the exception of taxa being
515 excluded from the optimized_taxa matrix coloured in dark grey (for details see Additional file 2:
516 Figure S1).

517 **Additional file 5: Table S2**. Overview of ephemeropteran mitochondrial genomes. Total sequence
518 length in base pairs (bp) and individual nucleotide compositions of mayfly mitochondrial genomes
519 calculated based on the whole available sequences and the ones corrected for the incomplete
520 mitochondrial genomes (see Methods).

521

522 **References**

523 1. Hovmöller R. The Palaeoptera Problem: Basal Pterygote Phylogeny Inferred from 18S and 28S
524 rDNA Sequences. Cladistics. 2002;18:313–23.

525 2. Ogden HT, Whiting MF. The problem with "the Paleoptera Problem:" sense and sensitivity.
526 Cladistics. 2003;19:432–42.

527 3. Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. Advances in Insect Phylogeny at
528 the Dawn of the Postgenomic Era. Annu. Rev. Entomol. 2012;57:449–68.

529 4. Whitfield JB, Kjer KM. Ancient Rapid Radiations of Insects: Challenges for Phylogenetic
530 Analysis. Annu. Rev. Entomol. 2008;53:449–72.

531 5. Ishiwata K, Sasaki G, Ogawa J, Miyata T, Su Z-H. Phylogenetic relationships among insect orders
532 based on three nuclear protein-coding gene sequences. Mol. Phylogenet. Evol. 2011;58:169–80.

533 6. Yeates DK, Cameron SL, Trautwein M. A view from the edge of the forest: recent progress in
534 understanding the relationships of the insect orders. Australian Journal of Entomology. 2012;51:79–
535 87.

536 7. Kjer KM, Simon C, Yavorskaya M, Beutel RG. Progress, pitfalls and parallel universes: a history
537 of insect phylogenetics. J R Soc Interface. 2016;13:20160363.

538 8. Mallatt J, Giribet G. Further use of nearly complete 28S and 18S rRNA genes to classify
539 Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 2006;40:772–94.

540 9. Meusemann K, Reumont von BM, Simon S, Roeding F, Strauss S, Kuck P, et al. A Phylogenomic
541 Approach to Resolve the Arthropod Tree of Life. Mol. Biol. Evol. 2010;27:2451–64.

542 10. Song N, Li H, Song F, Cai W. Molecular phylogeny of Polyneoptera (Insecta) inferred from
543 expanded mitogenomic data. Nature Publishing Group. 2016;6:36175.

544 11. Kjer KM, Carie FL, Litman J, Ware J. A Molecular Phylogeny of Hexapoda. Arthropod

545 Systematics & Phylogeny. 2006;64:35–44.

546 12. Thomas JA, Trueman JWH, Rambaut A, Welch JJ. Relaxed Phylogenetics and the Palaeoptera
547 Problem: Resolving Deep Ancestral Splits in the Insect Phylogeny. Syst. Biol. 2013;62:285–97.

548 13. Sasaki G, Ishiwata K, Machida R, Miyata T, Su Z-H. Molecular phylogenetic analyses support
549 the monophyly of Hexapoda and suggest the paraphyly of Entognatha. BMC Evol. Biol. BioMed
550 Central; 2013;13:236.

551 14. Reumont von BM, Jenner RA, Wills MA, Dell'Ampio E, Pass G, Ebersberger I, et al.
552 Pancrustacean Phylogeny in the Light of New Phylogenomic Data: Support for Remipedia as the
553 Possible Sister Group of Hexapoda. Mol. Biol. Evol. 2012;29:1031–45.

554 15. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, et al. Arthropod relationships
555 revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 2010;463:1079–83.

556 16. Ma C, Wang Y, Wu C, Kang L, Liu C. The compact mitochondrial genome of *Zorotypus*
557 *medoensis* provides insights into phylogenetic position of Zoraptera. BMC Genomics. 2014;15:1156.

558 17. Zhang J, Zhou C, Gai Y, Song D, Zhou K. The complete mitochondrial genome of Parafronurus
559 youi (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera. Gene. 2008;424:18–
560 24.

561 18. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC. The Strepsiptera problem: phylogeny of
562 the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and
563 morphology. Syst. Biol. 1997;46:1–68.

564 19. Giribet G, Ribera C. A Review of Arthropod Phylogeny: New Data Based on Ribosomal DNA
565 Sequences and Direct Character Optimization. Cladistics. 2000;16:204–31.

566 20. Ma Y, He K, Yu P, Yu D, Cheng X, Zhang J. The complete mitochondrial genomes of three
567 bristletails (Insecta: Archaeognatha): the paraphyly of Machilidae and insights into archaeognathan
568 phylogeny. Sun G, editor. PLoS One. 2015;10:e0117669.

569 21. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM, Carpenter. The Phylogeny of the Extant
570 Hexapod Orders. Cladistics. 2001;17:113–69.

571 22. Lin C-P, Chen M-Y, Huang J-P. The complete mitochondrial genome and phylogenomics of a
572 damselfly, *Euphaea formosa* support a basal Odonata within the Pterygota. Gene. 2010;468:20–9.

573 23. Wan X, Kim MI, Kim MJ, Kim I. Complete Mitochondrial Genome of the Free-Living Earwig,
574 *Challia fletcheri* (Dermaptera: Pygidicranidae) and Phylogeny of Polyneoptera. PLoS One.
575 2012;7:e42056.

576 24. Kjer K. Aligned 18S and Insect Phylogeny. Syst. Biol. 2004;53:506–14.

577 25. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A. Towards an 18S
578 phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed
579 nucleotide/doublet models. Zoology. 2007;110:409–29.

580 26. Reumont von BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, et
581 al. Can comprehensive background knowledge be incorporated into substitution models to improve

582 phylogenetic analyses? A case study on major arthropod relationships. BMC Evol. Biol. 2009;9:119.

583 27. Yoshizawa K, Johnson KP. Aligned 18S for Zoraptera (Insecta): phylogenetic position and
584 molecular evolution. Mol. Phylogenet. Evol. 2005;37:572–80.

585 28. Blanke A, Greve C, Wipfler B, Beutel RG, Holland BR, Misof B. The Identification of
586 Concerted Convergence in Insect Heads Corroborates Palaeoptera. Syst. Biol. 2013;62:250–63.

587 29. Simon S, Strauss S, Haeseler von A, Hadrys H. A Phylogenomic Approach to Resolve the Basal
588 Pterygote Divergence. Mol. Biol. Evol. 2009;26:2719–30.

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

591 31. Rokas A, Carroll SB. Bushes in the tree of life. PLoS Biol. Public Library of Science;
592 2006;4:e352.

593 32. Whitfield JB, Lockhart PJ. Deciphering ancient rapid radiations. Trends Ecol. Evol.
594 2007;22:258–65.

595 33. Talavera G, Vila R. What is the phylogenetic signal limit from mitogenomes? The reconciliation
596 between mitochondrial and nuclear data in the Insecta class phylogeny. BMC Evol. Biol.
597 2011;11:315.

598 34. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution.
599 Bioinformatics. 2005;21:2104–5.

600 35. Abascal F, Posada D, Zardoya R. MtArt: A New Model of Amino Acid Replacement for
601 Arthropoda. Mol. Biol. Evol. 2006;24:1–5.

602 36. Le VS, Dang CC, Le QS. Improved mitochondrial amino acid substitution models for metazoan
603 evolutionary studies. BMC Evol. Biol. BioMed Central; 2017;17:136.

604 37. Bensasson D. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol.
605 2001;16:314–21.

606 38. Rogers HH, Griffiths-Jones S. Mitochondrial Pseudogenes in the Nuclear Genomes of
607 Drosophila. PLoS One. 2012;7:e32593.

608 39. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding
609 overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc.
610 Natl. Acad. Sci. U.S.A. 2008;105:13486–91.

611 40. Claes KBM, De Leeneer K. Dealing with pseudogenes in molecular diagnostics in the next-
612 generation sequencing era. Methods Mol. Biol. New York, NY: Springer New York;
613 2014;1167:303–15.

614 41. Cameron SL. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny. Annu.
615 Rev. Entomol. 2014;59:95–117.

616 42. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999;27:1767–80.

617 43. Saito S. Replication Origin of Mitochondrial DNA in Insects. Genetics. 2005;171:1695–705.

618 44. Zhang D-X, Hewitt GM. Insect mitochondrial control region: A review of its structure, evolution
619 and usefulness in evolutionary studies. Biochem. Sys. Ecol. 1997;25:99–120.

620 45. Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM. Deducing the pattern of arthropod
621 phylogeny from mitochondrial DNA rearrangements. Nature. 1995;376:163–5.

622 46. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, et al.
623 Creating a buzz about insect genomes. Sills J, editor. Science. American Association for the
624 Advancement of Science; 2011;331:1386–6.

625 47. Barber-James HM, Gattolliat J-L, Sartori M, Hubbard MD. Global diversity of mayflies
626 (Ephemeroptera, Insecta) in freshwater. Hydrobiologia. Springer Netherlands; 2008;595:339–50.

627 48. Dijkstra K-DB, Bechly G, Bybee SM, Dow RA, Dumont HJ, Fleck G, et al. The classification
628 and diversity of dragonflies and damselflies (Odonata). In: Zhang Z-Q, editor. 2013. pp. 36–45.

629 49. Rutschmann S, Detering H, Simon S, Fredslund J, Monaghan MT. discomark: nuclear marker
630 discovery from orthologous sequences using draft genome data. Mol. Ecol. Resour. 2017;17:257–66.

631 50. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search
632 programs. Nucleic Acids Res. 1997;25:3389–402.

633 51. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
634 Bioinformatics. 2009;25:1754–60.

635 52. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
636 Bioinformatics. 2010;26:589–95.

637 53. Salmela L, Schroder J. Correcting errors in short reads by multiple alignments. Bioinformatics.
638 2011;27:1455–61.

639 54. Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. Evolution, Weighting, and
640 Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase
641 Chain Reaction Primers. Annals of the Entomological Society of America. Oxford University Press;
642 1994;87:651–701.

643 55. Li D, Qin J-C, Zhou C-F. The phylogeny of Ephemeroptera in Pterygota revealed by the
644 mitochondrial genome of *Siphluriscus chinensis* (Hexapoda: Insecta). Gene. 2014;545:132–40.

645 56. Aberer AJ, Krompass D, Stamatakis A. Pruning Rogue Taxa Improves Phylogenetic Accuracy:
646 An Efficient Algorithm and Webservice. Syst. Biol. 2012;62:162–6.

647 57. Wilkinson M. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol.
648 Evol. 1996;13:437–44.

649 58. Felsenstein J. Cases in which Parsimony or Compatibility Methods Will be Positively
650 Misleading. Systematic Zoology. 1978;27:401.

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

653 60. Landan G, Graur D. Local reliability measures from sets of co-optimal multiple sequence
654 alignments. Pac Symp Biocomput. 2008;:15–24.

655 61. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable
656 alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res.
657 2015;43:W7–14.

658 62. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
659 Improvements in Performance and Usability. Mol. Biol. Evol. 2013;30:772–80.

660 63. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: Combined Selection of Partitioning
661 Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol. Evol. 2012;29:1695–701.

662 64. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
663 phylogenies. Bioinformatics. 2014;30:1312–3.

664 65. Bergsten J. A review of long-branch attraction. Cladistics. 2005;21:163–93.

665 66. Kück P, Struck TH. BaCoCa – A heuristic software tool for the parallel assessment of sequence
666 biases in hundreds of gene and taxon partitions. Mol. Phylogenet. Evol. 2014;70:94–8.

667 67. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2:
668 Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst.
669 Biol. 2012;61:539–42.

670 68. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language.
671 Bioinformatics. 2004;20:289–90.

672 69. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.

673 70. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and
674 annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol.
675 2017;8:28–36.

676 71. R Development Core Team. R: A language and environment for statistical computing. R
677 Foundation for Statistical Computing, Vienna, Austria. International Statistical Review. 2012. pp.
678 357–70.

679 72. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved de
680 novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 2013;69:313–9.

681 73. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA.
682 Bioinformatics. 2004;20:3252–5.

683 74. Laslett D, Canback B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial
684 nucleotide sequences. Bioinformatics. 2008;24:172–5.

685 75. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in
686 genomic sequence. Nucleic Acids Res. Oxford University Press; 1997;25:955–64.

687 76. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW--a suite of tools for
688 generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets.

689 Nucleic Acids Res. 2013;41:W575–81.

690 77. Perna N, Kocher T. Patterns of nucleotide composition at fourfold degenerate sites of animal
691 mitochondrial genomes. J Mol Evol. 1995;41.

692 78. Rota-Stabelli O, Yang Z, Telford MJ. MtZoa: a general mitochondrial amino acid substitutions
693 model for animal evolutionary studies. Mol. Phylogenet. Evol. 2009;52:268–72.

694 79. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the
695 timing and pattern of insect evolution. Science. 2014;346:763–7.

696 80. Simon S, Narechania A, DeSalle R, Hadrys H. Insect Phylogenomics: Exploring the Source of
697 Incongruence Using New Transcriptomic Data. Genome Biol. Evol. 2012;4:1295–309.

698 81. Ogden TH, Whiting MF. Phylogeny of Ephemeroptera (mayflies) based on molecular evidence.
699 Mol. Phylogenet. Evol. 2005;37:625–43.

700 82. Ogden TH, Gattolliat JL, Sartori M, Staniczek AH, Soldán T, Whiting MF. Towards a new
701 paradigm in mayfly phylogeny (Ephemeroptera): combined analysis of morphological and molecular
702 data. Sys. Entomol. 2009;34:616–34.

703 83. Letsch HO, Meusemann K, Wipfler B, Schutte K, Beutel R, Misof B. Insect phylogenomics:
704 results, problems and the impact of matrix composition. Proc. R. Soc. B. 2012;279:3282–90.

705 84. Letsch H, Simon S. Insect phylogenomics: new insights on the relationships of lower neopteran
706 orders (Polyneoptera). Sys. Entomol. 2013;38:783–93.

707 85. Zhang Y-Y, Xuan W-J, Zhao J-L, Zhu C-D, Jiang G-F. The complete mitochondrial genome of
708 the cockroach *Eupolyphaga sinensis* (Blattaria: Polyphagidae) and the phylogenetic relationships
709 within the Dictyoptera. Mol. Biol. Rep. 2009;37:3509–16.

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

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

715 88. Inward D, Beccaloni G, Eggleton P. Death of an order: a comprehensive molecular phylogenetic
716 study confirms that termites are eusocial cockroaches. Biology Letters. 2007;3:331–5.

717 89. Legendre F, Nel A, Svenson GJ, Robillard T, Pellens R, Grandcolas P. Phylogeny of
718 Dictyoptera: Dating the Origin of Cockroaches, Praying Mantises and Termites with Molecular Data
719 and Controlled Fossil Evidence. PLoS One. 2015;10:e0130127.

720 90. Regier JC, Mitter C, Zwick A, Bazinet AL, Cummings MP, Kawahara AY, et al. A large-scale,
721 higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies).
722 Moreau CS, editor. PLoS One. 2013;8:e58568.

723 91. Wu H-Y, Ji X-Y, Yu W-W, Du Y-Z. Complete mitochondrial genome of the stonefly
724 *Cryptoperla stilifera* Sivec (Plecoptera: Peltoperlidae) and the phylogeny of Polyneopteran insects.
725 Gene. 2014;537:177–83.

726 92. Pons J, Bauzà-Ribot MM, Jaume D, Juan C. Next-generation sequencing, phylogenetic signal
727 and comparative mitogenomic analyses in Metacrangonyctidae (Amphipoda: Crustacea). BMC
728 Genomics. 2014;15:566.

729 93. Tang M, Tan M, Meng G, Yang S, Su X, Liu S, et al. Multiplex sequencing of pooled
730 mitochondrial genomes--a crucial step toward biodiversity analysis using mito-metagenomics.
731 Nucleic Acids Res. 2014;42:e166–6.

732 94. Lee EM, Hong MY, Kim MI, Kim MJ, Park HC, Kim KY, et al. The complete mitogenome
733 sequences of the palaeopteran insects *Ephemera orientalis* (Ephemeroptera: Ephemeridae) and
734 *Davidius lunatus* (Odonata: Gomphidae). Genome. 2009;52:810–7.

735 95. Kim MJ, Jung KS, Park NS, Wan X, Kim K-G, Jun J, et al. Molecular phylogeny of the higher
736 taxa of Odonata (Insecta) inferred from COI, 16S rRNA, 28S rRNA, and EF1-α sequences.
737 Entomological Research. 2014;44:65–79.

738