Molecular Phylogenetics and Evolution 55 (2010) 846–859

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Ribosomal protein genes of holometabolan insects reject the Halteria, instead revealing a close affinity of Strepsiptera with Coleoptera

Stuart J. Longhorn [a,b,*], Hans W. Pohl [c], Alfried P. Vogler [a,b]

[a] Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK
[b] Division of Biology, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK
[c] Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum, Friedrich-Schiller-Universität Jena, D-07743 Jena, Germany

ARTICLE INFO

Article history:
Received 9 June 2009
Revised 18 February 2010
Accepted 19 March 2010
Available online 27 March 2010

Keywords:
Ribosomal proteins
ESTs
Holometabola
Strepsiptera
Composition
Stemminess/RCV
Topological congruence
Long-branch attraction

ABSTRACT

The phylogenetic relationships among holometabolan insect orders remain poorly known, despite a wealth of previous studies. In particular, past attempts to clarify the sister-group of the enigmatic order Strepsiptera with rRNA genes have led to intense debate about long-branch attraction (the 'Strepsiptera problem'), without resolving the taxonomic question at hand. Here, we appealed to alternative nuclear sequences of 27 ribosomal proteins (RPs) to generate a data matrix of 10,731 nucleotides for 22 holometabolan taxa, including two strepsipteran species. Phylogenetic relationships among holometabolan insects were analyzed under several nucleotide-coding schemes to explore differences in signal and systematic biases. Saturation and compositional bias particularly affected third positions, which greatly differed in AT content (18–72%). Such confounding factors were best reduced by R-Y coding and removal of third codon positions, resulting in more strongly supported topologies, whereas amino acid coding gave poor resolution. The placement of Strepsiptera with Coleoptera (the Coleopterida) was recovered under most coding schemes and analytical methods, if often with modest support and ambiguity. In contrast, an alternative sister-group with Diptera (the Halteria) was only found in one analysis using parsimony, and weakly supported. The topologies here generally support a Coleoptera + Strepsiptera as sister-group to Mecopterida (Siphonaptera + Mecoptera + Diptera + Lepidoptera + Trichoptera), while Hymenoptera were always recovered as sister-group to the remaining Holometabola.

1. Introduction

Insects constitute a huge proportion of Earth's biodiversity, both in terms of described species and biomass (Grimaldi and Engel, 2005). Furthermore, the history of insect diversification provides an unparalleled example of massive speciation and extensive specialization. Partly due to the immense scale of insect biodiversity, the relationships among several orders remain uncertain, despite radical new approaches to taxonomy and systematics (Beutel and Pohl, 2006; Hennig, 1953, 1981; Kristensen, 1995, 1999; Whiting, 2002b). Most extant insects (about 80% of described species) undergo complete metamorphosis, and hence are placed together as Holometabola (=Endopterygota). This major group is itself dominated by four hyper-diverse orders, Coleoptera (beetles), Lepidoptera (butterflies and moths), Hymenoptera (bees, wasps and ants), and Diptera (flies, mosquitoes), with 38%, 16%, 13% and 13% of insect species described to date.

There is little consensus on the phylogenetic relationships among holometabolan orders from morphological data (Kristensen, 1991, 1999) (Fig. 1A and B). The most widely accepted supra-ordinal group is Amphiesmenoptera, which unites Trichoptera (caddisflies) and Lepidoptera (Kristensen, 1997). The other well-accepted higher grouping is Neuropterida, which unites Megaloptera (alderflies, etc.), Neuroptera (lacewings) and Raphidioptera (snakefles), but which together may only warrant ordinal status (Aspöck, 2002; Hörnschemeyer, 2002). Most supra-ordinal groups are not well supported. Several authors have suggested that Diptera (true flies), Mecoptera (scorpionflies), and Siphonaptera (fleas, which may fall within 'Mecoptera'; see below) should be united as Antliophora (Beutel and Pohl, 2006; Hennig, 1953, 1981; Hörnschemeyer, 2002; Kristensen, 1991, 1999; Wootton, 2002). More broadly, the Antliophora and Amphiesmenoptera have been tentatively united as the Mecopterida (elsewhere called Panorpida) (Fig. 1B).

Perhaps the most controversial question for insect systematics is the uncertain position of the enigmatic order Strepsiptera (twisted-wing parasites) (reviewed in Beutel and Pohl, 2006; Kristensen, 1999). The Strepsiptera are a cosmopolitan order of obligate endoparasites with only about 600 species worldwide.

* Corresponding author at: Dept. of Biology, National University of Ireland, Maynooth, EIRE, Ireland.
E-mail addresses: sjl197@hotmail.com (S.J. Longhorn), hans.pohl@uni-jena.de (H.W. Pohl), a.vogler@nhm.ac.uk (A.P. Vogler).

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.03.024

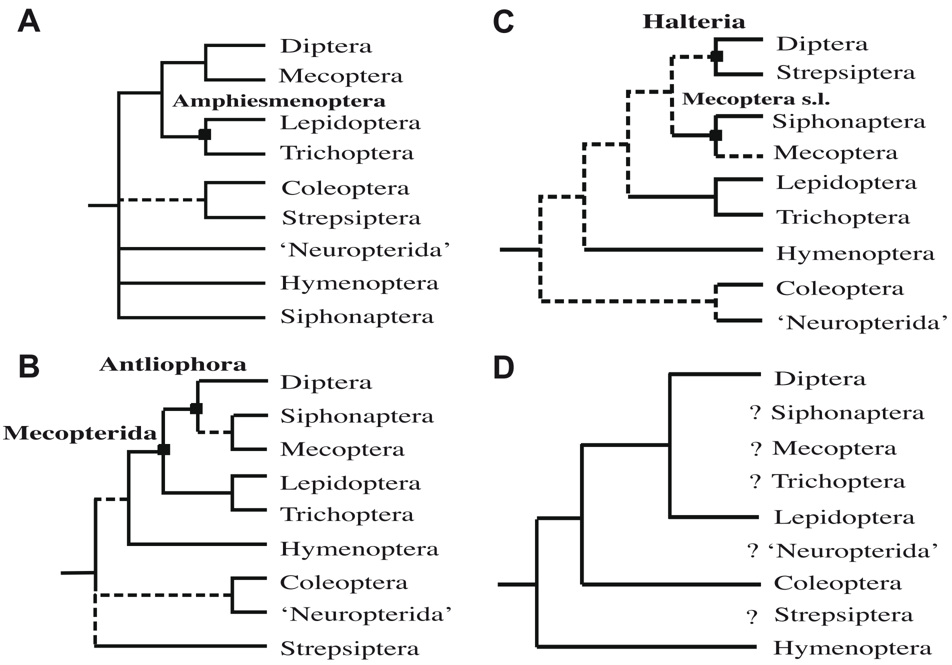

Fig. 1. Hypothetical relationships of holometabolan insect orders from various types of data. (A) Morphology (Hennig, 1981); (B) morphology (Kristensen, 1991); (C), 18S + 28S rRNA + morphology (Whiting, 2002a,b; Whiting, 2004); (D), phylogenomic data of 185 genes from completed genomes (Savard et al., 2006). Dashed branches represent doubtful monophyly or weak support identified by the original authors, while question marks refer to unsampled lineages. Key: 'Neuropterida' = Megaloptera + Neuroptera + Raphidioptera. 'Mecoptera' refers to Mecoptera *sensu lato* (*i.e.* paraphyletic "Mecoptera" with Neomecoptera sister to Siphonaptera).

Several morphological studies have proposed that Strepsiptera are allied to Coleoptera (Beutel and Gorb, 2001; Kathirithamby, 1989; Kinzelbach and Pohl, 2003; Kukalová-Peck and Lawrence, 1993) or, more radically, are highly-modified derived Coleoptera allied to the polyphagan family Rhipiphoridae, which also parasitize Hymenoptera and undergo hypermetamorphosis (Crowson, 1960). In general reviews, Strepsiptera are typically placed in an unresolved position among Holometabola (*e.g.* Kristensen, 1999; Grimaldi and Engel, 2005), whilst even their status as holometabolan insects has been questioned (Beutel and Pohl, 2006).

Initial molecular studies using ribosomal RNA (rRNA) provided strong support for an unexpected sister-group between Strepsiptera and Diptera (Chalwatzis et al., 1995; Whiting and Wheeler, 1994) (Fig. 1C and D). This contentious group was named 'Halteria' to suggest these orders could be united morphologically by the presence of wings modified to halteres (Whiting and Wheeler, 1994), despite the fact that modifications affect the forewings in Strepsipteran and hindwings in Diptera. Other characters proposed to support the close affinities of these two orders (*e.g.* Wheeler et al., 2001; Whiting et al., 1997) have since been critically reinterpreted (Beutel and Pohl, 2006; Hünefeld and Beutel, 2005; Krenn, 2007; Kristensen, 1999; Pohl et al., 2005; Sinclair et al., 2007). Studies using 28S rRNA data conflicted with results from 18S rRNA alone, as they did not recover the Halteria, and even placed Strepsiptera outside Holometabola altogether (Huelsenbeck, 2001; Hwang et al., 1998). One major complication is that insect rRNA sequences are affected by extensive rate heterogeneity and compositional differences (Friedrich and Tautz, 1997; Gillespie et al., 2005; Huelsenbeck, 2001; Hwang et al., 1998), so that misleading signal can challenge phylogenetic methods. Such complications have led to the total exclusion of Strepsiptera from some more recent phylogenetic analyses of insect relationships with rRNA (Kjer, 2004; Kjer et al., 2006; Whitfield and Kjer, 2008). Other aspects of the holometabolan phylogeny have also been modified by insights from rRNA, for example indicating the placement Siphonaptera within Mecoptera. This renders the latter paraphyletic (Wheeler et al., 2001; Whiting et al., 1997), which has since been supported when combined with morphology (Whiting, 2002a, 2004) (Fig. 1C).

Molecular studies using nuclear genes other than rRNA have consistently failed to recover the contentious Strepsiptera–Diptera sister-group (Halteria). For example, the strepsipteran *engrailed* homeobox gene does not possess a derived intron shared by Diptera and Lepidoptera, but which is also absent from Coleoptera or Hymenoptera, and hence does not provide a synapomorphy for Halteria (Rokas et al., 1999). Similarly, the ecdysone receptor (UPS/RXR) shares a conserved functional motif in Lepidoptera and Diptera, which is absent in Strepsiptera and several other orders (Hayward et al., 2005; Bonneton et al., 2006). Both unique features may otherwise be interpreted as possible synapomorphies for the wider Mecopterida, to the exclusion of Strepsiptera. The most robust evidence against the Halteria has been the recent phylogenetic analysis of six single-copy genes, which instead grouped Strepsiptera as sister-group to Coleoptera (Wiegmann et al., 2009). However, taxon sampling of Coleoptera was limited to two species of Chrysomelidae and Tenebrionidae, representing closely related families in the derived suborder Polyphaga, rather than a preferred broad sampling across the four coleopteran suborders (Hughes et al., 2006; Hunt et al., 2007). This leaves much uncertainty about the Strepsiptera–Coleoptera affiliation, and in particular about possible confounding branch attraction between these lineages, which may be reduced by a comprehensive taxon sampling of Coleoptera. Elsewhere in this six-gene analysis, taxon sampling is more representative of lineage diversity. Overall, their findings do not support previous controversial results from rRNA based datasets, such as recovering Mecoptera as the monophyletic sister-group to Siphonaptera within a broader Antliophora and Mecopterida (Wiegmann et al., 2009).

Recent genome and Expressed Sequence Tag (EST) sequencing from diverse insects now provides new possibilities for the analysis of basal relationships in Holometabola, without need for targeted PCR amplifications (Hughes et al., 2006; Roeding et al., 2007; Savard et al., 2006). Whole genome sequences from Coleoptera (*Tribolium*), Diptera (*Anopheles*, *Aedes*, *Drosophila* spp.), Hymenoptera (*Apis*) and Lepidoptera (*Bombyx*) have already revealed a novel placement of Hymenoptera as sister-group to all remaining holometabolan orders (Krauss et al., 2008; Savard et al., 2006; Zdobnov and Bork, 2007) (Fig. 1D). In addition, large EST libraries for insects of socioeconomic importance are available, including disease vectors (*Anopheles*, *Aedes*, *Glossina*, *Pediculus*, *Ctenocephalides*), agricultural pests (*Tribolium*, *Toxoptera*, *Manduca*, *Bombyx*), food producers (*Apis*) and genetic models (*Spodoptera*, *Drosophila*). Complemented with EST sequences from a suite of insects selected for taxonomic breadth (Hughes et al., 2006), the available data provide a wide representation of holometabolan insect phylogenetic diversity.

Here, we compile nuclear ribosomal protein (RP) gene sequences from existing and newly obtained EST data and genome sequences for a broad sample of holometabolan orders and selected outgroups. RPs are commonly detected in EST libraries due to their high transcription rates, and have several desirable features as phylogenetic markers (Hughes et al., 2006; Theodorides et al., 2002) including low paralogy (Landais et al., 2003), synergistic phylogenetic signal (Longhorn et al., 2007) and wide genomic distribution at unlinked sites (Marygold et al., 2007). We re-evaluated holometabolan insect relationships with focus on the contentious position of the Strepsiptera, using newly curated EST data to expand taxon sampling beyond previous studies. Various coding schemes and tree building methods applied to this RP dataset do not show support for the Strepsiptera–Diptera sister-group of previous rRNA based-studies (*i.e.* the Halteria hypothesis), but favor instead the affinity of Strepsiptera with Coleoptera.

2. Materials and methods

2.1. cDNA libraries of Strepsiptera and RP detection

Our gene sampling was designed to extract all available RP sequences from published EST libraries of two Strepsiptera, *Eoxenos laboulbenei* and *Mengenilla chobauti* (both family Mengenillidae), followed by searches for the same genes for a broad sample of other Holometabola. Strepsipteran data were obtained from sequencing of cDNA libraries (forward and reverse sequencing of clones with >600 base inserts) to produce 346 and 302 high quality unigenes for *Eoxenos* and *Mengenilla*, respectively (Genbank CD492361–492706 and CD485197–485498), after exclusion of mitochondrial and rRNA transcripts (Hughes et al., 2006). The core 79 RPs from *Spodoptera frugiperda* (Landais et al., 2003) were used as queries for BLAST searches against strepsipteran ESTs in a local database, resulting in detection of 27 transcripts with high similarity to RP genes, specifically 15 RP genes in *Eoxenos* and 17 in *Mengenilla*. For each RP, redundant ESTs were clustered and ambiguities edited with Sequencher (Gene Codes Corp) using the original chromatograms (Hughes et al., 2006), which generally resulted in full-length transcripts. A similar process of data searching and editing was performed on published ESTs from Coleoptera (except *Tribolium* and *Ips*) given in Hughes et al. (2006) of wide taxonomic disparity, from among which the selection of taxa with the most complete gene-representation promoted matrix completion. With broad taxon sampling, we were able to include data from 3 out of 4 recognized coleopteran suborders, from *Sphaerius* (Myxophaga), *Cicindela* (Adephaga) and five diverse others (all suborder Polyphaga). In the exceptional case of *Cicindela* sp. we combined

Table 1
Source of arthropod RPs ranked by total EST.

Species	Taxonomy	ESTs	RPs	%[a]	Tissue	Dir.[b]	Source[c]
Drosophila melanogaster	Dip: Brac: Drosophilidae	382,439	27	96.8	Various	Both	TIGR
Aedes aegypti	Dip: Nem: Culicidae	171,414	27	99.9	Various	Both	TIGR
Anopheles gambiae	Dip: Nem: Culicidae	150,042	27	98.6	Various	Both	TIGR
Bombyx mori	Lep: Bombycidae	116,541	27	100	Various	Both	dbEST
Apis mellifera	Hym: Apidae	22,652	25	88.8	Neural	Unk.	dbEST
Acyrthosiphon pisum	Hem: Aphididae	22,183	27	96.8	Whole	Fwd.	dbEST
Glossina morsitans	Dip: Brac: Glossinidae	21,427	17	65.3	Midgut	Both	dbEST
Spodoptera frugiperda	Lep: Noctuidae	5937	27	97.9	Ovarian Sf9	Unk.	Nr.1[d]
Ctenocephalides felis	Sip: Pulicidae	4841	24	81.7	Hidgut	Fwd.	dbEST
Toxoptera citricida	Hem: Aphididae	4307	25	86.5	Whole	Fwd.	dbEST
Culicoides sonorensis	Dip: Nem: Ceratopogonidae	2979	22	79.7	Sal[f]. Midgut	Fwd.	dbEST
Tribolium castaneum	Col: Poly: Tenebrionidae	2465	26	97.2	Various	Both	dbEST
Plutella xylostella	Lep: Plutellidae	2131	27	94.0	Larvae	Unk.	dbEST[e]
Manduca sexta	Lep: Sphingidae	2009	24	85.6	Antennae	Fwd.	dbEST
Ips pini	Col: Poly: Scolytidae	1671	13	47.9	Midgut	Both	dbEST
Cicindela sp. (3 libraries)	Col: Ade: Cicindelidae	1386	13	43.6	Whole/testes	Both	OUR
Pediculus humanus	Phi: Pediculicidae	1127	15	55.8	Whole	Fwd.	dbEST
Georissus sp.	Col: Poly: Georissidae	892	9	28.0	Whole	Both	OUR
Agriotes lineatus	Col: Poly: Elateridae	771	11	35.5	Whole	Both	OUR
Papilio dardanus	Lep: Papilionidae	699	16	62.4	Wing Disc	Both	OUR
Sphaerius sp.	Col: Myx. Sphaeriusidae	696	10	29.8	Whole	Both	OUR
Timarcha balearica	Col: Poly: Chrysomelidae	658	10	33.3	Whole	Both	OUR
Tricolepisma aurea	Zygentoma: Lepismatidae	425	15	56.4	Whole	Fwd.	OUR
Eoxenos laboulbenei	Str: Megenillidae	346	17	60.2	Whole	Fwd.	OUR
Mengenilla chobauti	Str: Megenillidae	302	15	47.9	Whole	Fwd.	OUR
Hydropsyche sp.	Tri: Hydropsychidae	225	11	38.2	Silk Gland	Fwd.	dbEST

The data for *Cicindela* sp. was formed from 3 EST libraries of closely related taxa, with 8 RPs from *C. campestris*, 4 from *C. littoralis*, and 1 from *C. litorea*. Abbreviations: Adep [Adephaga]; Brac [Brachycera]; Col [Coleoptera]; Nem ['Nematocera']; Dip [Diptera]; Hem [Hemiptera]; Hym [Hymenoptera]; Lep [Lepidoptera]; Myx = [Myxophaga]; Phi [Phthiraptera]; Poly [Polyphaga]; Sip [Siphonaptera]; Str [Strepsiptera]; Tri [Trichoptera].

[a] Percentage of total matrix, after Gblocks exclusion of uncertain homology.

[b] Direction Sequenced; Fwd = N-terminus (5′) sequenced; Both = two directions (5′ and 3′) sequenced.

[c] TIGR = Gene Indicies, TIGR., dbEST and Nr. = GenBank, OUR = From our lab, available on GenBank or on request.

[d] Non-redundant (nr) database of GenBank together with unpublished ESTs (I. Landais, pers. Comm.).

[e] As unpublished data from J.H. Eum, Pers. Comm.

[f] Sal = Salivary gland.

data from EST libraries of three closely related species (see Table 1).

Our search for publicly available ESTs resulted in a database for the 27 target RPs from a further 24 insect species (Table 1). BLAST was used to identify redundant transcripts for each query taxon, which were subsequently clustered in Sequencher. A concern with using EST data is the low quality reads, possibly resulting in error rates of up to 3% (Sorek and Safer, 2003). However, this was ameliorated by careful re-assessment of the original chromatograms and removal of poor quality reads. Low-level variability between ESTs, which was limited to a small subset of sequences, was removed by clustering redundant ESTs into majority rule consensus transcripts. This approach also concealed some heterogeneity from transcript variants (alleles, recent paralogs, expressed pseudogenes) by incorporating any such minor variations into consensus sequences, especially when data redundancy was high. Non-target sequences, like xenologs, were easily removed during clustering. For example, ESTs from the tsetse fly *Glossina* frequently contained divergent RP transcripts that further investigation showed were contaminations from the protozoan *Trypanosoma.* Similarly, distant paralogs, such as mitochondrial RPs were also easily identified and excluded, especially once motifs in un-translated regions were considered. Overall, there was little evidence for functional duplications in our 27 RP gene set, though divergent homologous transcripts could occasionally be ascribed to specific lineages alone, mostly in *Drosophila*, and presumably resulted from recent gene duplications. In these cases, we again used consensus transcripts to mask the influence that variable sites of alternative copies might have on analyses.

2.2. Multiple sequence alignments

Recent multi-gene studies typically rely solely on amino acid characters (Hughes et al., 2006; Roeding et al., 2007; Savard et al., 2006). Accordingly, we investigated the utility of this RP dataset at the amino acid level, but additionally appealed to the information content of the underlying nucleotide sequences. Taxa were included only if at least 9 of 27 genes were available, a level chosen to allow broadest taxon sampling while restricting missing data. In particular, this threshold allowed us to include a wide diversity of Coleoptera, plus maximise the proportion of sites shared between them. Orthologous proteins were aligned as amino acids using ClustalX (Thompson et al., 1997). We then removed regions of alignment ambiguity with Gblocks 0.91b (Castresana, 2000), excluding all gapped sites (option -b5), but otherwise using default settings. The removed sections largely corresponded to regions with low similarity across taxa and/or containing gaps, both factors that can generate low confidence in homology assignments. Initial protein alignments were co-aligned to the DNA sequences with CodonAlign 1.0 (B.G. Hall, unpublished), from which alignment-ambiguous portions were then also removed to match the reduced protein alignments, *i.e.* removing the same regions identified by Gblocks in protein guide-alignments. Any linked histone and ubiquitin-like fusion proteins were also discarded, retaining only true RP segments for the analyses.

2.3. Phylogenetic analyses of nucleotides and recoding schemes

Selection for translational efficiency in highly expressed transcripts like RPs can complicate phylogenetic analyses of nucleotides due to strong compositional preferences and differential site evolution (heterogeneity), particularly at the deepest nodes. We therefore applied recoding and data partitioning strategies and compared phylogenetic signal at inter-ordinal nodes of interest, aiming to maximise internal over terminal branch length (see below, tree ‘stemminess’), and ameliorate misleading influences of site heterogeneity. For nucleotide data, $1_{NT}2_{NT}3_{NT}$ refers to a matrix retaining all three codon positions each with standard base coding, while exclusion of 3rd positions gave the reduced matrix $1_{NT}2_{NT}$, again with standard coding. We also used R-Y (purine–pyrimidine) coding, *e.g.* in the $1_{NT}2_{NT}3_{RY}$ data where third positions were R-Y recoded.

Tree searches on nucleotide sequences were performed with partitioned Bayesian inference, maximum likelihood (ML) and parsimony analysis. Bayesian analyses were conducted with MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001). We used the GTR+I+Γ_4 model (nst = 6), which was the best-fit according to the Akaike information criteria (AIC) in ModelTest 3.7 (Posada and Crandall, 1998). The same GTR model was favored when models were assessed for each codon position independently, but here each partition differed in transition/transversion ratios, proportion of invariable sites and among site rate variation. To account for site heterogeneity in Bayesian analyses, all model parameters (statefreq, revmat, shape, pinvar) were unlinked across codon partitions and variable rate multipliers used (prset ratepr = variable). A Dirichlet distribution was used for the rate matrix and base frequencies, under a prior of all tree topologies to be equally probable. For sites subject to R-Y coding, we set nst = 1 according to the CF87 + I + Γ_4 model, to specifically model transversions. For all datasets, we ran two concurrent Bayesian analyses of 2×10^6 generations with four chains (three heated and one cold) with different random starting trees (nruns = 2), and sampled every 500 generations. We examined the posterior probability distribution in Tracer 1.2 (Rambaut and Drummond, 2003) to determine stationarity, and confirmed the average deviation of split frequencies from the two runs was below 0.005 to verify convergence. From each dataset, we discarded the first 500 samplings as burn-in.

Initial model parameters for ML searches were selected as for Bayesian analyses. Searches were conducted in PAUP∗ 4.0b10 (Swofford, 2002) by successive searches of increasing intensity, with one Nearest Neighbor Interchange (NNI) replicate, two rounds of Subtree Pruning and Regrafting (SPR) and two Tree-Bisection-Reconnection (TBR) searches, until the log likelihood converged across search replicates. A global molecular clock was also evaluated in PAUP under these parameters, and likelihood ratio tests conducted with all branch lengths assumed to be equal. Branch support for ML analyses were calculated from 1000 non-parametric bootstrap replicates with PhyML 2.4.4 (Guindon and Gascuel, 2003), except the $1_{RY}2_{RY}$ dataset where we ran only 100 bootstrap replicates in PAUP∗ due to computational limitations. Model parameters for ML searches were the same as described for Bayesian analyses above. Parsimony analyses were conducted with all sites weighted equally, with heuristic searches and TBR branch swapping on 100 random-addition replicates. Branch support was assessed using PAUP∗ from 1000 bootstrap replicates.

2.4. Phylogenetic analyses of amino acids

For analyses of amino acid characters, we conducted likelihood (with PhyML) and Bayesian analyses (MrBayes) using the Dayhoff model with empirical residue frequencies, gamma rate distribution and invariant sites, as suggested by both AIC and BIC tests in ProtTest 1.2.6 (Abascal et al., 2005). Both tests actually indicated the LG substitution model as the best-fit, but this is not implemented in MrBayes (and failed to reach convergence with Phylobayes, see below). Therefore we resorted to the Dayhoff model for initial MrBayes analyses, which was preferred by ProtTest over other alternatives such as WAG and JTT. To explore effects of alternative amino acid substitution models, we also conducted additional Bayesian analyses with an unconstrained model prior in MrBayes (using prset aamodelpr = mixed). Here, the JTT model was strongly favored with high posterior probability (of 1.00). Finally, we ran additional Bayesian analyses in Phylobayes 3.1g (Lartillot et al., 2009) with the preferred LG model (-lg -ratecat), or with free exchangeabilities among amino acid states (-gtr -cat), allowing relative rates to be considered free parameters (CAT-GTR). We also conducted parsimony searches (PAUP) on amino acids, where nodal support was assessed using the bootstrap as for nucleotide analyses.

2.5. Assessment of signal to noise ratio

We evaluated the phylogenetic signal using stemminess (Fiala and Sokal, 1985) and relative compositional variability (RCV) (Phillips et al., 2004, 2006). These measures were used to determine if altering the data-coding scheme might improve the ratio of phylogenetic signal against stochastic noise. Stemminess describes the ratio of internal to terminal branch lengths. Stemminess was calculated as the sum of internal branch lengths divided by the sum of all branch lengths in the tree, *i.e.* the proportion of tree distance attributed to internal branches. High stemminess results if a greater proportion of character changes are attributed to internal branches relative to those on terminal branches. Assuming that multiple hits have accumulated at the ('older') internal branches while terminal branches are less affected, short terminal branches (relative to internal branches) indicate a general low rate of reversals throughout the tree. In contrast, a predominance of changes on terminal branches (extending their length relative to internal branches) means that even the least susceptible recent branches are already affected by multiple hits, and deep branches even more so. Hence, if the removal or recoding of certain characters in the matrix improves the overall stemminess in the resulting topologies, this is interpreted here to indicate that saturation and potential long-branch attraction artifacts are reduced. Stemminess values were calculated on uncorrected neighbor-joining trees generated in PAUP, all constrained to the same topology, with uninformative (constant and autapomorphic) sites excluded (Phillips et al., 2006). Nucleotide heterogeneity under different coding schemes was assessed using RCV with uninformative sites removed, and is defined as the average variability of composition between taxa, calculated from n taxa and t sites across the matrix (Phillips et al., 2004). The ratio of stemminess/RCV (Phillips et al., 2006; Yamanoue et al., 2007) is then interpreted to reflect the degree to which multiple hits may be affected by directional nucleotide compositional bias, *i.e.* is used as a measure of the potential for long-branch attraction across the entire data set.

2.6. Testing alternative topologies

To test competing hypothesis of holometabolan ordinal relationships, we conducted likelihood searches with various fixed topological constraints, enforcing selected nodes or nested sets. The constrained nodes were Coleoptera [*Agriotes*, *Cicindela*, *Georissus*, *Ips*, *Sphaerius*, *TiMarcha*, *Tribolium*], Halteria = Diptera [*Aedes*, *Anopheles*, *Culicoides*, *Drosophila*, *Glossina*] + Strepsiptera [*Eoxenos*, *Mengenilla*], Paraneoptera = Phthriaptera [*Pediculus*] + Hemiptera [*Acrythosiphon*, *Toxoptera*], Antliophora = Diptera [see above] + Siphonaptera [*Ctenocephalides*], and Mecopterida = Diptera + Siphonaptera + Lepidoptera [*Bombyx, Manduca*, *Papilio*, *Plutella*, *Spodoptera*] + Trichoptera [*Hydropsyche*]. For constrained analyses with Antliophora or Mecopterida, we accepted Siphonaptera as surrogate for the unsampled Mecoptera (*sensu* Whiting, 2002a,b) due to evidence that Siphonaptera are either a derived mecopteran lineage (Whiting, 2002a) or the sister-group (Wiegmann et al., 2009).

We evaluated the likelihood of each constrained topology using both approximately unbiased (AU) tests and weighted Shimodaira–Hasegawa (wSH) tests, and used both tests that differ in their intrinsic statistical properties to clarify the confidence around *P*-value estimates. The AU test (Shimodaira, 2002) was used for general hypothesis testing in Consel (Shimodaira and Hasegawa, 2001) with the multi-scale bootstrap (P-boot). First, site-wise log likelihoods for the ML tree and constraint topologies were generated with PAUP∗. We then generated several sets of bootstrap replicates of varying sequence length, and calculated *P*-values by comparing differences in bootstrap support among different length sets for the hypothesis at hand. The AU test provides a means of assessing the confidence of tree selection, and unlike other tests overcomes the problem of selection bias from simultaneous comparisons of multiple trees, which can lead to overconfidence in erroneous trees. We also employed SH tests (Shimodaira and Hasegawa, 1999) for multiple topological comparisons, which are preferred over related Kishino–Hasegawa tests due to their greater suitability for comparisons between an *a posteriori* topology (*i.e.* the topology with highest likelihood) against alternative test scenarios. In addition, SH tests are generally less sensitive to rejection of alternative hypotheses (type II error) than AU tests, and hence give a more conservative perspective on the plausibility of alternative scenarios with the data (Shimodaira, 2002). Here, we adopted the weighted implementation (wSH) to reduce the problem of over-conservation, by using the maximum of the standardized difference of log

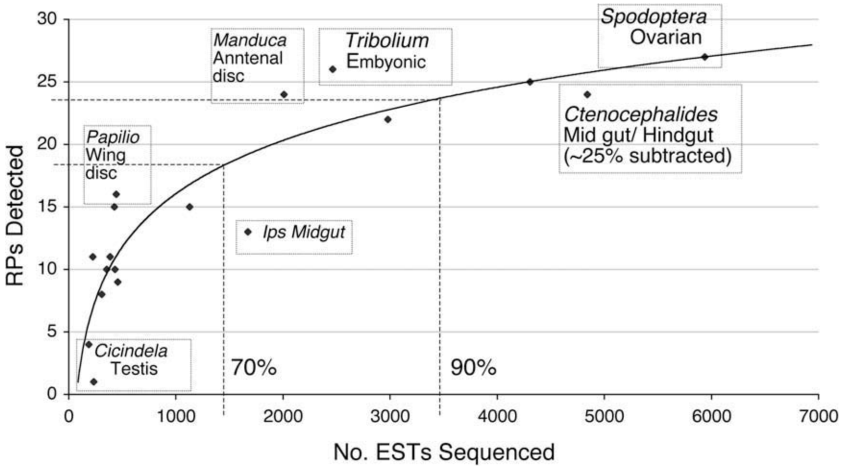

Fig. 2. Detection rates of the 27 RP genes across insects with <7000 ESTs (at time of sampling) from N-terminus ESTs alone, excluding non-independent reverse reads. These results fitted a logarithmic trend (6.144 Ln(x), $r^2 = 0.827$, Pearson Coefficient $P < 0.01$). Large libraries (>7000 ESTs) were excluded as complete sets of 27 genes were commonly detected.

likelihoods as the test statistic (Buckley et al., 2001). Furthermore, the weighted version also corrects for tree selection bias. We also conducted wSH tests in Consel, using the RELL approximation for bootstrap values to avoid costly re-calculation of parameters.

3. Results

3.1. Detection and conservation of RP genes

Taxon sampling for the 27 target RP genes present in two Strepsiptera libraries focused on holometabolan insects, but includes selected Hemiptera and Phthiraptera from the outgroup Paraneoptera to evaluate the potential placement of Strepsiptera outside Holometabola. Full genome sequences and/or >100,000 ESTs were available for several Diptera (*Drosophila*, *Aedes*, *Anopheles*), Lepidoptera (*Bombyx*), Hymenoptera (*Apis*), Coleoptera (*Tribolium*) and the paraneopteran Hemiptera (*Acrythosiphon*). For these taxa, the matrix of 27 RP genes is largely complete. These taxa also show greatest data redundancy, often with several hundred sequences per gene (Suppl. Table 1). At the extremes, *Drosophila melanogaster* has the highest mean transcript redundancy of 249 ESTs per gene, followed by *Anopheles gambiae* with a mean of 218 ESTs. For taxa with lower matrix completion, including most Coleoptera, several Lepidoptera and both Strepsiptera, transcript redundancy is much lower, typically with only one or two ESTs per gene. Across all taxa, there is a significant logarithmic correlation between gene detection and the number of sampled ESTs (Fig. 2). This suggests that our complete 27 RP set is commonly detected when library size is around 7000 ESTs. In contrast, RP transcript detection is below average in taxa with small or normalized EST libraries or from constitutive tissues like gut (*Ctenocephalides*, *Ips*), but well above average in embryonic tissue (*Tribolium*) or developmentally active tissues like imaginal discs (*Papilio*, *Manduca*, *Spodoptera*).

After exclusion of ambiguously aligned sites identified by Gblocks, the reduced matrix comprises 3577 aligned amino acids (10,731 nucleotides) with 68% completion (Table 1), and an average of 18.5 genes per taxon. The proportion of unambiguously aligned sites for each taxon ranges from 28% to 100%, with an average of 65% sites sampled. If we consider only variable sites, data completion is reduced to an average of 46% (Suppl. Table 2). The two Strepsiptera *Eoxenos* and *Mengenilla* share an average of 38% and 36% variable sites with the other terminals in the matrix, with 60% and 48% completion respectively, but combined are nearly 100% complete. The lowest density of sampled sites between taxon pairs is within Coleoptera (excluding *Tribolium*), at an average of 26% of matrix completion at variable sites. However, *Tribolium* provided a dense scaffold of RP data here with 97% completion, and shared on average 36% of variable sites with other Coleoptera, and from 30% to 98% of variable sites with other taxa in the matrix.

3.2. The phylogeny of Holometabola as inferred from RP sequences

The full dataset including all nucleotides ($1_{NT}2_{NT}3_{NT}$ data) revealed great differences in nucleotide composition (Suppl. Fig. 1), as indicated by the high RCV of 0.41 (Table 2) and significant deviation from homogeneity with the Chi-square test of homogeneity in PAUP ($\chi^2 = 3625.4$, $P = 0.00$). Although all codon positions varied in AT composition, this was much greater in the third codon positions, ranging from 79% AT in *Apis* (Hymenoptera) to only 18% in *Drosophila* (Diptera) (Suppl. Table 3). Phylogenetic trees from parsimony analysis (Suppl. Fig. 1, right) resulted in many unexpected groupings, which appear to be driven by these compositional differences. Taxa with highest AT-richness tended group together (most notably *Apis* with *Pediculus*, *Glossina and Culicoides*), while very AT poor taxa were placed together elsewhere (*i.e. Drosophila*, *Anophleles* and *Aedes*, with some Lepidoptera and others). In contrast, ML and Bayesian analyses (Suppl. Fig. 1, left) both appear to be more consistent with several expected ordinal nodes including, for example, the monophyly of Diptera was recovered with high nodal support despite their widely different AT compositions. In model-based analyses, the two Strepsiptera grouped together with high nodal support, and were placed together with most Coleoptera, either in Bayesian analyses as sister-group to *Cicindela* (suborder Adephaga), or in ML analyses as sister-group to a wider group of Coleoptera from the two major suborders Adephaga + Polyphaga.

Exclusion of third positions ($1_{NT}2_{NT}$ data) increases the proportion of invariable sites from 41% to 61% (Table 2), reduces the number of parsimony informative sites from 52% to 30%, and increases stemminess from 0.15 in the $1_{NT}2_{NT}3_{NT}$ dataset to 0.30. Removal of third positions also reduces compositional heterogeneity, with RCV falling from 0.41 to 0.35, although the composition remains significantly different across taxa in this reduced data (all sites: $\chi^2 = 250.3$, $P = 0.00$; informative sites only: $\chi^2 = 654.9$, $P = 0.00$). Reduced compositional differences through exclusion of third positions also leads to a decrease in potential long-branch attraction, with the ratio of internal branches/heterogeneity (stemminess/RCV) rising from 0.38 to 0.87, presumably due to a greater influence of retained sites with slower evolutionary rates and reduction in saturated sites.

The reduced $1_{NT}2_{NT}$ data (Fig. 3, left) groups the two Strepsiptera together with high support and places them within a poorly resolved Coleoptera, which are rendered paraphyletic. With Bayesian analysis, Strepsiptera is weakly supported as sister-group to the coleopteran suborders Adephaga (*Cicindela*) + Polyphaga (five taxa). ML analysis places Strepsiptera as sister-groups to a monophyletic Coleoptera, while parsimony suggests these affinities are unresolved. There is strong support for other key nodes, including the monophyly of Lepidoptera and its sister-group relationship with Trichoptera (=Amphiesmenoptera). Dipteran monophyly is

Table 2
Stemminess and RCV values for alternative subsets of the 27 RP dataset.

Data partition[a]	Sites	Constant	Parsimony informative	Stemminess[b]	RCV[c]	Stemminess[c]/RCV
$1_{NT}2_{NT}3_{NT}$	10,731	4440 (41%)	5550 (52%)	0.154 (0.156)	0.409	0.38
$1_{NT}2_{NT}3_{RY}$	10,731	5509 (51%)	4295 (40%)	0.254 (0.257)	0.342	0.74
$1_{NT}2_{NT}$	7154	4353 (61%)	2139 (30%)	0.306 (0.307)	0.352	0.87
$1_{RY}2_{NT}$	7154	4823 (67%)	1706 (24%)	0.378 (0.387)	0.345	1.10
$1_{RY}2_{RY}$	7154	5181 (72%)	1451 (20%)	0.346 (0.351)	0.341	1.02
2_{NT}	3577	2485 (69%)	754 (21%)	0.393 (0.394)	0.341	1.15
Amino acid	3577	2012 (56%)	1129 (32%)	N/a	N/a	N/a

[a] $1_{NT}2_{NT}3_{NT}$, all positions included (subscript "NT" denotes nucleotides); $1_{NT}2_{NT}3_{RY}$, with third codon positions recoded into purine (R)–pyrimidine (Y) states (subscript "RY" denotes R-Y coding); etc.

[b] Values with negative branch lengths allowed (in brackets, negative branch lengths constrained to zero), all calculated without uninformative sites, *i.e.* no constant or autapomorphic sites.

[c] Calculated first with codon positions separated, then summarized.

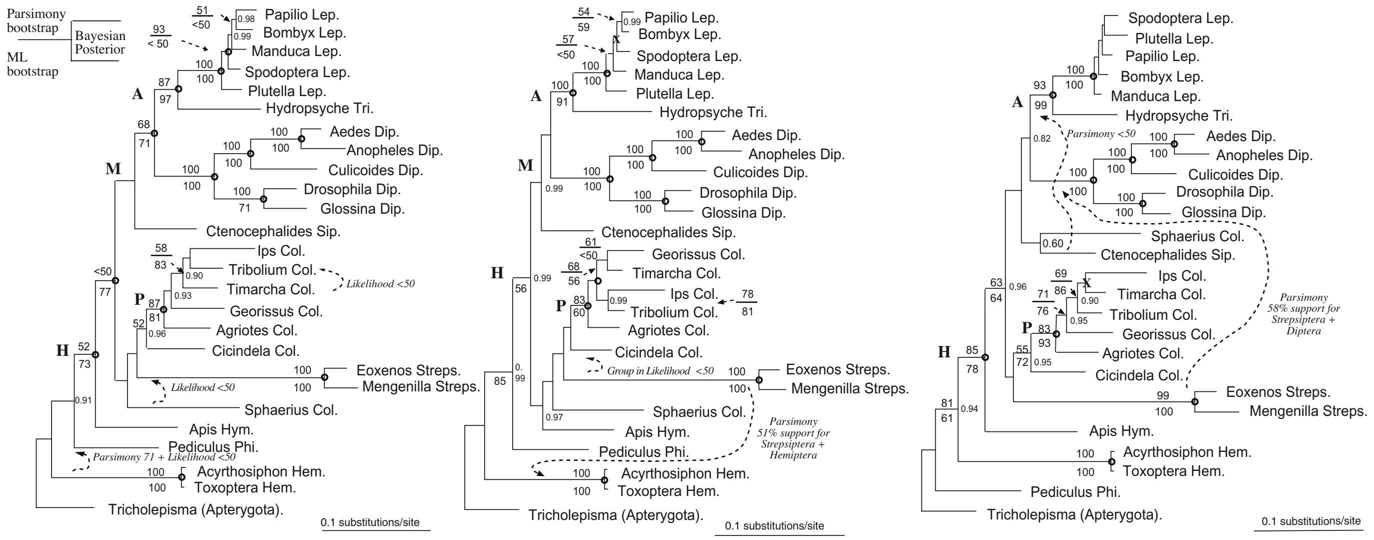

Fig. 3. Partitioned Bayesian analyses of the 27 RP genes under alternate nucleotide recoding schemes. Left, standard second codon positions alone [2_{NT}]. Centre, first and second positions with standard coding [$1_{NT}2_{NT}$]. Right, first and second positions together both R-Y-coded [$1_{RY}2_{RY}$]. Posterior probabilities inside nodes, with those of absolute support (PP = 1.0) indicated by circles on nodes. Bootstrap scores from likelihood are above branches, and bootstrap from parsimony below. Here, A = Amphiesmenoptera (Trichoptera + Lepidoptera), M = Mecopterida (Diptera + Amphiesmenoptera + Siphonaptera [+Mecoptera]), P = Polyphaga Coleoptera, and H = Holometabola (Inc. Strepsiptera).

also strongly supported, with the separation of 'nematoceran' and brachyceran lineages clearly evident, though long internal branch lengths compared to other holometabolan orders. At deeper nodes, we obtain moderate bootstrap support and high posterior probability for a novel sister-group relationship of Diptera and Amphiesmenoptera (=Lepidoptera + Trichoptera) rather than Antliophora (=Diptera + Siphonaptera [+unsampled Mecoptera]). Bayesian analysis strongly supports the monophyly of the Holometabola including Strepsiptera, with Hymenoptera as sister-group to all remaining Holometabolan orders, as recently found by other multi-gene studies.

When second positions are used alone (the 2_{NT} data), the proportion of constant sites increases slightly again, and stemminess also rises from 0.31 to 0.39. In addition, RCV is further reduced so that compositional differences are now non-significant ($\chi^2 = 63.5$, $P = 0.82$, after excluding uninformative sites), and the signal to noise ratio (stemminess/RCV) increases to 1.15, the highest value of all data sets investigated (Table 2). Yet, while the resultant topologies are similar to the larger $1_{NT}2_{NT}$ data, branch support is generally weak at internal nodes (Fig. 3, centre). Again, with Bayesian and ML approaches, Strepsiptera is placed within a paraphyletic Coleoptera, albeit weakly supported. In contrast, parsimony analysis places Strepsiptera as sister-group to the outgroup Hemiptera, *i.e.* outside of the Holometabola, but with poor bootstrap support. Even though the 2_{NT} dataset is compositionally homogeneous, both the strepsipteran and hemipteran lineages show increased branch lengths compared to most other groups, so more likely influenced by branch attraction artifacts. If a molecular clock is enforced with a GTR model, so that equal rates are assumed across lineages, the null hypothesis of clocklike rates is rejected at high significance ($P < 0.001$, likelihood ratio test) indicating that rate differences may indeed have affected the tree topology.

Next, we conducted analyses with R-Y coding (see Methods) to ameliorate the misleading influence of saturation and compositional bias. With R-Y coding at both first and second positions, higher stemminess and lower RCV come at the expense of approximately one third fewer characters compared to standard coding of the same sized $1_{NT}2_{NT}$ data (Table 2). Now, compositional differences are non-significant ($\chi^2 = 18.3$, $P = 0.83$, after excluding uninformative sites). As before, Bayesian and ML analyses recover Strepsiptera as the sister-group of Coleoptera, but parsimony analysis places Strepsiptera with Diptera (Fig. 3, right), yielding the only case to recover the Halteria. However, nodal support for either placement is low. An unconventional grouping of *Sphaerius* (Coleoptera, Myxophaga) and *Ctenocephalides* (Siphonaptera) is recovered under model-based methods, though also associated with weakly supported nodes, and perhaps related to loss of variation in this homogeneous and we suspect unnecessarily conservative $1_{RY}2_{RY}$ dataset.

As a compromise to reduce heterogeneity, yet to retain a maximum of signal, we adopted an intermediate $1_{RY}2_{NT}$ scheme, which retains a greater number of parsimony informative sites, while stemminess at 0.38 is the highest overall except for the smaller 2_{NT} (Table 2). RCV is also comparably low and compositional heterogeneity is non-significant in the Chi-square test ($\chi^2 = 96.1$, $P = 0.51$), reflecting the beneficial reduction of heterogeneity with R-Y coding. The stemminess/RCV ratio at 1.10 was amongst the highest values for all matrices. These scores suggest the mixed $1_{RY}2_{NT}$ coding scheme is the best balance between large matrix size, homogeneous composition, noise reduction, and maximal tree-length at internal branches. The resulting topologies are well resolved (Fig. 4, left), with support for several supra-ordinal relationships higher than under other coding schemes (compare Fig. 3). Strepsiptera again cluster with Coleoptera, either as sister-group to *Cicindela* (suborder Adephaga) under parsimony, as sister-group to Polyphaga + Adephaga with Bayesian analyses, or as sister-group to a monophyletic Coleoptera with ML analyses. Most other nodes are identical to results from other coding schemes, such as strong support for Diptera and several expected sub-lineages (including the monophyly of culicoid mosquitoes). Again, there is strong support for a sister-group between Lepidoptera and Trichoptera (Amphiesmenoptera). We also now find moderate support for the Mecopterida (Amphiesmenoptera plus Diptera and Siphonaptera), but again, the position of Siphonaptera is inconsistent with Antliophora (Diptera + Siphonaptera/Mecoptera group), though the alternative Diptera + Amphiesmenoptera is only weakly supported. Overall, the monophyly of Holometabola including Strepsiptera is well supported, with Hymenoptera as sister-group to the remaining Holometabola. The lack of monophyly

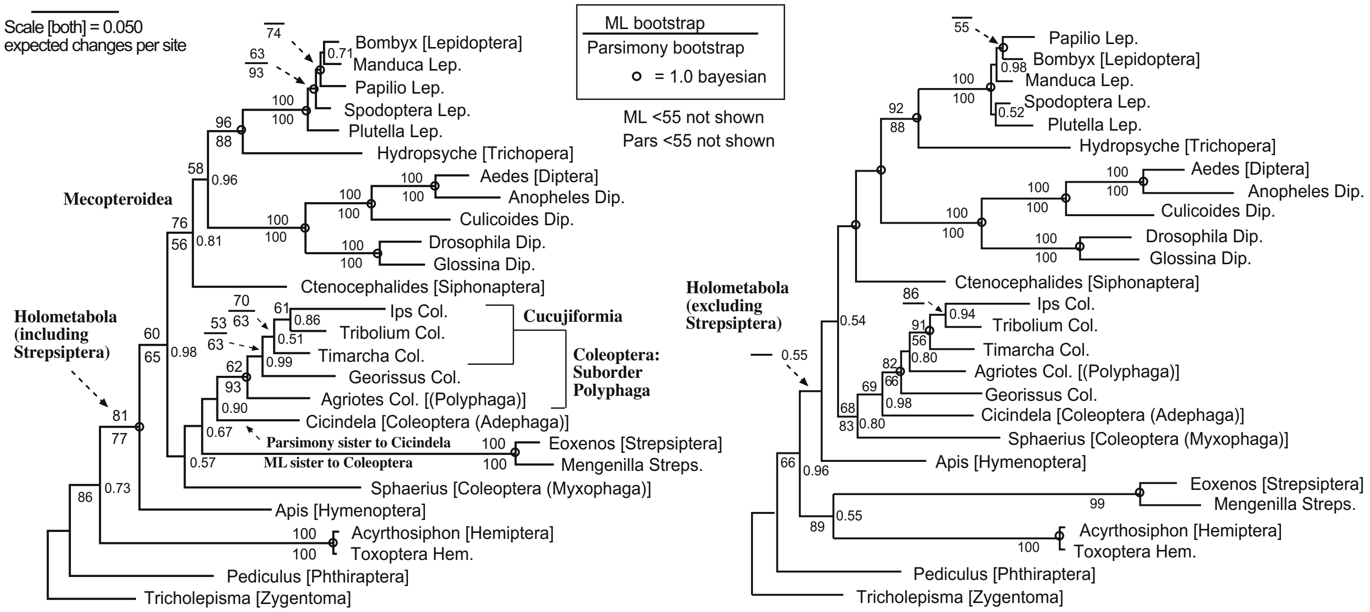

Fig. 4. Revised Bayesian analyses for the 27 RPs. Left, from 7154 nucleotides using partitioned analysis of the $1_{RY}2_{NT}$ scheme. Parsimony gave two solutions (5108 steps; C.I. = 0.46; R.I. = 0.59) that only disagreed with an unresolved trichotomy for polyphagan Coleoptera (*Ips*, *TiMarcha* and *Tribolium*). Right, Bayesian analysis from 3577 amino acids using the Dayhoff model. The ML topology was identical (−Ln = 34718.09698), as was the single solution from parsimony (4391 steps; 1129 informative characters; C.I. = 0.64; R.I. = 0.66). Branch support annotations as in Fig. 3.

for Paraneoptera [Hemiptera + Phthiraptera] is surprising, but might be explained at an artifact of long-branch attraction to the outgroup, as apterygote insects (*Tricholepisma*) provide a rather distant root.

Phylogenetic analyses of amino acids (Fig. 4, right) give similar results to nucleotide data, except for a surprising location of Strepsiptera as sister to Hemiptera (Paraneoptera), which is well supported in parsimony (BS = 89%), but not model-based analyses (likelihood BS < 50%, Bayesian PP = 0.55). Unlike the use of nucleotide sequences, the corresponding amino acids provide moderate nodal support for the monophyly of Coleoptera, with the basal suborder Myxophaga as sister-group to Adephaga plus Polyphaga. However, low branch support is typical at most supra-ordinal nodes. Amphiesmenoptera clusters with Diptera, supported by high Bayesian posteriors, but not supported by ML or parsimony bootstrap. Here, the position of Siphonaptera conflicts with the Antliophora, but again recovers the wider Mecopterida.

Further phylogenetic analyses of amino acids also tend to place Strepsiptera and Coleoptera together. In Bayesian analyses without a fixed amino acid model prior (allowing model jumping with aamodelpr = mixed in MrBayes), the JTT model was strongly preferred (Fig. 5), yielding an identical topology to trees from the $1_{RY}2_{NT}$ dataset. Here, most ordinal nodes received high posterior support (*e.g.* Lepidoptera, Diptera, etc.), as does the supra-ordinal group Amphiesmenoptera, but there is only weak support for monophyletic Coleoptera (PP = 0.80) and most constituent nodes, except for a well supported Polyphaga, while Strepsiptera is sister-group to Coleoptera (PP = 0.61). Alternative amino acid analyses in Phylobayes with a fixed LG model failed to reach convergence after a run of 500 h on a high-specification bioinformatics cluster, and so must be treated with caution (Suppl. Fig. 2). As with the Dayhoff model (Fig. 4 right), Strepsiptera are placed with Hemiptera with moderate support (PP = 0.80), although both on long-branches, and associated with weaker support for monophyly of the remaining Holometabola (PP = 0.75). The other nodes are consistent with previous analyses, and Coleoptera are monophyletic. Finally, when Bayesian analyses of amino acids are conducted with free relative exchangeabilities among states (using CAT-GTR), the topology is almost identical to the mixed model results (Suppl. Fig. 3 versus Fig. 5), though with Strepsiptera inside a paraphyletic Coleoptera, as sister-group to *Sphaerius* (Coleoptera, Myxophaga) (PP = 0.56), but with good posterior support for a wider Coleoptera plus Strepsiptera group (PP = 0.98). Under this method, there was also good support for the monophyly of Holometabola including Strepsiptera (PP = 0.92).

3.3. Testing the stability of alternative evolutionary scenarios

Hypothesis testing on a nested series of tree topologies was then conducted to establish stability of phylogenetic scenarios (Table 3) using the $1_{RY}2_{NT}$ dataset, which shows the greatest support at supra-ordinal nodes, and hence most topological stability. These analyses reveal that coleopteran monophyly (constraint 2) is not rejected at the 5% significance level with either AU or wSH tests ($P = 0.79$ and $P = 0.98$, respectively), hence this represents a plausible alternative topology to the ML tree. Likewise, Paraneoptera monophyly, for the traditional placement of aphids (Hemiptera) with the louse *Pediculus* (Phthiraptera) is also not rejected at the 5% level, either enforced on its own (constraint 3, $P = 0.13$ and $P = 0.41$) or together with the monophyly of Coleoptera (constraint 4, $P = 0.17$ and $P = 0.49$). The Antliophora (Diptera + Siphonaptera; constraint 5) is also not rejected ($P = 0.11$ and $P = 0.32$). Under all these constraints, ML searches repeatedly recover the preferred trees with Strepsiptera either as sister-group to Coleoptera, or Coleoptera except *Sphaerius* (Suborder Myxophaga).

We also specifically tested alternative placements of Strepsiptera. When Strepsiptera are enforced as sister-group to Diptera (constraint 9; *i.e.* the Halteria hypothesis) the topology of highest likelihood has the Diptera–Strepsiptera pair as sister-group to Amphiesmenoptera (Lepidoptera + Trichoptera), then Siphonaptera outside, contradicting Antliophora. Here, the AU result and the more conservative wSH test are $P = 0.06$ and $P = 0.18$, respectively, *i.e.* close to rejection of this topology. An alternative position of Strepsiptera within the broader Mecopterida (constraint 6) is not rejected by either AU or wSH tests ($P = 0.23$ and $P = 0.47$), but under these conditions the topology of highest likelihood has

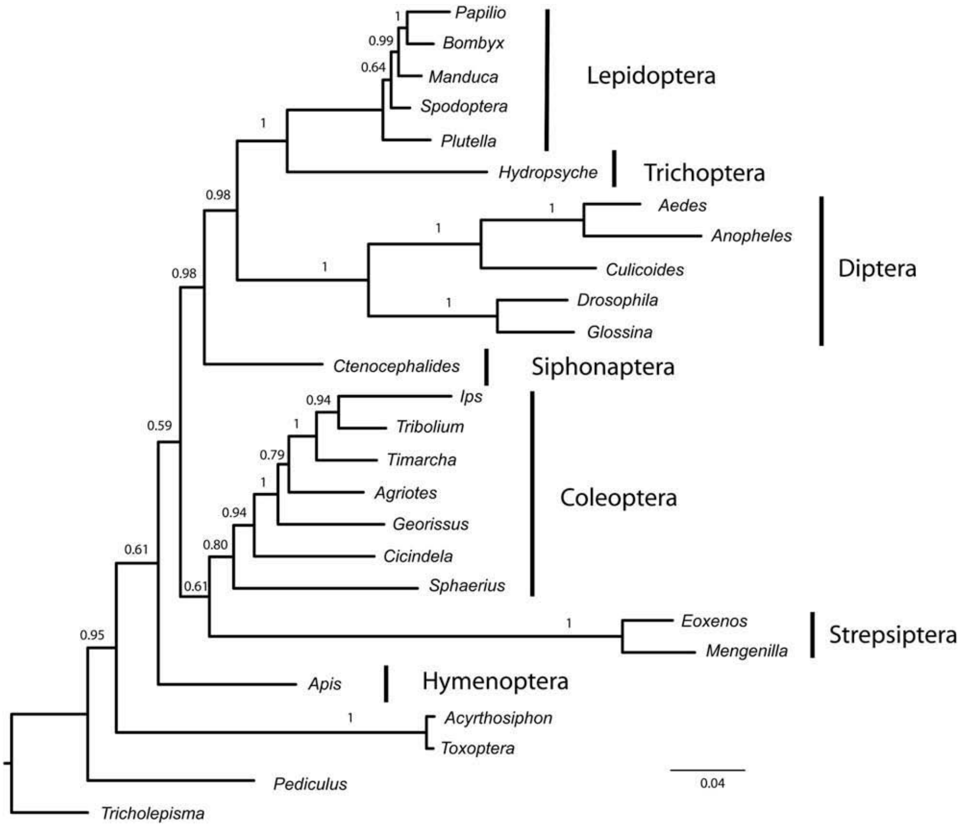

Fig. 5. Bayesian analysis of amino acids. A mixed model prior was applied, and the search included two chains of two million generations (sample-freq = 500, burn-in = 1001), leading to a sampling of 6000 trees and average deviation of split frequencies of 0.02 indicating convergence, and PSRF = 1 for TL, alpha and pinvar. In both duplicate runs, 0.00% of the chain accepted changes were due to alterations in the model (parameter 6), and the JTT model received a posterior probability of 1.00 (scale = expected changes per site).

Table 3
Concordance tests of alternative topologies using the preferred $1_{RY}2_{NT}$ dataset.

Topology	Strepsiptera sister to?	Ln Score	δ	AU test	wSH test	P-Boot
1: Unconstrained [MaxLike]	Coleoptera except *Sphaerius* [Myxophaga]	30136.4598	–	0.732	0.938	0.420
2: Coleoptera	Coleoptera	30136.5458	0.086	0.790	0.975	0.374
3: Paraneoptera	Coleoptera except Sphaerius	30140.3303	3.871	0.137	0.411	0.048
4: Paraneoptera, Coleoptera	Coleoptera	30140.4468	3.987	0.171	0.490	0.040
5: Antliophora	Coleoptera	30142.4571	5.997	0.106	0.324	0.034
6: Mecopterida* [=Inc. Strepsiptera]	Remaining 'Mecopterida'	30142.8687	6.409	0.229	0.472	0.036
7: Holometabola except Hymenoptera	Coleoptera + Mecopterida [*i.e.* most Holometabola]	30143.6062	7.146	0.039	0.307	0.004
8: Antliophora + Paraneoptera	Coleoptera	30146.3736	9.914	0.038	0.152	0.004
9: Halteria* [=Inc. Strepsiptera]	Diptera [+unconventional Amphiesmenoptera]	30154.9232	18.463	0.061	0.184	0.025
10: Holometabola except Strepsiptera [with Coleoptera + Hymenoptera]	Holometabola	30159.2348	22.775	0.055	0.093	0.014
11: Halteria + Antliophora* [=Inc. Strepsiptera]	Diptera [+ conventional Siphonaptera]	30161.4670	25.007	0.005	0.061	2e-04
12: Holometabola except Strepsiptera [Hymenoptera basal in remaining]	Holometabola	30161.5469	25.087	0.010	0.028	0.001
13: Holometabola except Strepsiptera + Antliophora [*i.e.* exc. Strepsiptera]	Holometabola	30164.5145	28.055	0.012	0.048	0.001
14: Halteria + Antliophora* [=Inc. Strepsiptera + Paraneoptera	Diptera [then together sister to Siphonaptera]	30165.3934	28.934	4e-04	0.035	0
15: Antliophora, Mecopterida* [=Inc. Strepsiptera]	Antliophora [all inside Mecopterida]	30165.4851	29.025	3e-03	0.016	0

See methods for details of taxa included in each constraint, but when name marked by asterix, the constraint refers to the supra-ordinal clade name, but includes Strepsiptera in the constraint.

Strepsiptera sister-group to all remaining 'mecopteridan' orders, *i.e.* not closely related to Diptera. Conversely, our constraint for the traditional interpretation of the holometabolan phylogeny from 18S rRNA data (constraint 11) is rejected ($P = 0.005$, AU test; 0.06 wSH test). Here, Strepsiptera are constrained as sister-group to Diptera (*i.e.* Halteria) both within constrained Antliophora* for

((Diptera + Strepsiptera) Siphonaptera) and the wider Mecopterida*. This is also strongly rejected if Paraneoptera (Hemiptera + Phthiraptera) is concurrently enforced (constraint 14; $P = 3e\ -03$, 0.03). When Strepsiptera are constrained within Mecopterida*, but now sister-group to Antliophora (constraint 15) as ((Diptera + Siphonaptera) Strepsiptera), the topology is even more strongly rejected with both tests ($P = 3e\ -03$, $P = 0.01$). All other positions of Strepsiptera, including as sister-group to all holometabolan orders are rejected in at least one of the tests, with phylogenetic scenarios of Strepsiptera as the basal branch in the Holometabola (constraints 12, 13) among the most strongly rejected (Table 3).

4. Discussion

4.1. RP dataset generation and possible limitations

The 27 nuclear RPs provide a rich new source of to address major relationships in Holometabola, including the position of the Strepsiptera, studied here with EST derived data for the first time. The taxonomic representation of RP data in currently available EST libraries and genome sequences was highly appropriate to test the two main hypotheses for the position of Strepsiptera; as sister-group to either Diptera (the Halteria hypothesis) or Coleoptera. Both potential sister-groups were represented by a broad range of internal lineages, although insufficient data were available for the Mecoptera (as close relatives to Diptera), nor from Neuroptera, Raphidioptera, and Megaloptera (the Neuropterida, as close relatives to Coleoptera). A major effort of the current study went towards the careful curation of the dataset; ESTs are usually deposited as raw sequence data and phylogenetic analyses potentially suffer from poor base calling (besides enzyme errors in the reverse transcription step). This is rarely problematic for taxa with large EST libraries, where majority consensus can be used to produce reliable sequences of transcripts with few or no ambiguities. For the smaller libraries, including most Coleoptera and the Strepsiptera analyzed here, careful re-examination of chromatograms was necessary to produce the final high quality reads from fewer individual cDNA clones which, unlike most EST data, were sequenced in both directions to improve read quality and facilitated recovery of full-length transcripts.

The 27 full-length RP genes obtained from the two Strepsiptera, provided a scaffold for the compilation of data from other insect groups. Both strepsipteran taxa are members of the family Mengenillidae and therefore for purposes of establishing higher-level relationships in the Holometabola could be combined as a single chimeric taxon in future analyses. This would also be justified by the fact that both taxa always grouped together with high support, indicating the uniform signal from both (and the power of the few overlapping sites to recover their monophyly). For the majority of our analyses Strepsiptera clustered with Coleoptera, but without strong support. There is no reason to suspect the data missing in our matrices had a persuasive impact on the phylogenetic placement of the Strepsiptera, given that at least one of the two possible strepsipteran sequences were gathered from all 27 RP genes. Further, other species with complete or near-complete gene sampling had a broad taxonomic distribution, so our matrices were largely complete at the ordinal level. The broad taxonomic sample of Coleoptera with three of four suborders, and five major polyphagan lineages also strengthens the utility of the data set. Here, broad taxon sampling was achieved by incorporating several smaller-scale EST libraries (Table 1). Although matrix completion was relatively low for most Coleoptera with an average of 25.6% shared sites from 10 to 13 genes, the gene set for *Tribolium* was largely complete. We generally obtained good resolution within Coleoptera, specifically strong support for the suborder Polyphaga, and internal relationships largely in agreement with other studies (Hughes et al., 2006; Hunt et al., 2007).

The general recovery of expected relationships within Coleoptera and high nodal support for the monophyly of the two Strepsiptera together suggest that despite substantial missing data in these taxa, the remaining sites contained enough signal for phylogenetic inferences. This is also indicated, for example, in strong support for the well-accepted Amphiesmenoptera despite substantial missing data in Trichoptera (*i.e.* only 38% matrix completion for *Hydropsyche* sp., average 28% variable shared sites with any other taxa). Other multi-gene studies indicate the missing data do not necessarily affect the recovered topology where numerous positions are still available for sufficient signal, although resolution and nodal support may be reduced with respect to comparable matrices of greater completion (Bapteste et al., 2002; Wiens, 2003; Philippe et al., 2004). Consequently, we see no reason to suspect that elevated matrix completion would imply any different topology, allaying concerns that missing data may have confounded the phylogenetic placement of the Strepsiptera here. While differences in matrix completion may have influenced levels of nodal support, we see little reason to suspect that missing data *per se* produced erroneous phylogenetic signal in these analyses.

4.2. Alternative coding schemes to overcome systematic bias in nucleotide variation

A more pervasive problem is the inference of deep level relationships affected by high nucleotide variation. Multi-gene analyses of this kind (and larger phylogenomic studies) have almost invariably resorted to data transformation by protein translation, thereby reducing some problems of frequent substitutions that lead to saturation and systematic biases (*e.g.* Bapteste et al., 2002; Philippe et al., 2004). However, this practice does not use the fullest extent of primary sequence information, losing potential signal from synonymous nucleotide changes. Whereas high rates and multiple changes at a site *per se* are not misleading the tree, and high homoplasy over the entire topology may still be informative within subclades (Källersjö et al., 1999), the pressing concern is about the convergent biases in sequence evolution that may confound phylogenetic inference, in particular with an increased level of sequence divergence. In the current data set, taxon-specific differences in nucleotide composition were exceptionally large, and much greater than in mtDNA data sets for which the uniform use of all nucleotides has been advocated widely (Källersjö et al., 1999; Wenzel and Siddall, 1999; Vogler et al., 2005). Uniform coding of all sites at the nucleotide level resulted in trees that appeared to contain spurious nodes, in particular in the parsimony analyses that contradicted monophyly of well-established orders such as the Diptera (Suppl. Fig. 1 and Table 3). Erroneous grouping appears to be driven by base composition, *e.g.* for the flies, where the highly AT rich (*Culicoides*, *Glossina*) and GC rich (*Aedes*, *Anopheles*, *Drosophila*) taxa are misplaced with other insects of similar nucleotide composition, instead of together as expected for the well-established order Diptera. The nucleotide bias affects predominantly third positions and presumed silent sites in first positions (Suppl. Table 3). The removal or recoding of these sites provides a test of their confounding effect, as the recoding strategies applied here and removal of the most variable sites together provide complementary methods to ameliorate the influence of compositional heterogeneity and transition bias (Phillips et al., 2004).

The combined effects of saturation and compositional biases were mainly assessed using 'tree stemminess' and RCV. Although there can be other reasons for low stemminess, *e.g.* if short internal branches reflect a fast radiation early in a lineage, our results are consistent with saturation. Tree stemminess greatly increased on removal of the most variable third codon positions, while R-Y recoding of remaining sites to retain just the rarer transversions further increased tree stemminess (Table 2). As rare changes result in a higher ratio of the relative length of internal to terminal branches, this indicates that changes in the fast-varying sites are underestimated. Such saturation appears to predominantly affect the third codon positions, the removal of which results in doubling the stemminess score, while transitions in first positions, but not in second positions, are seemingly also affected by saturation.

In addition, we assessed the RCV, which provides a measure of the deviation of nucleotide composition in each taxon from the mean composition, and which therefore is useful to assess nucleotide bias across the matrix. Compositional bias can be expected to strongly affect characters most likely to undergo multiple substitutions, especially the synonymous changes in the third and first codon positions, and taxa that have experienced character convergence may be incorrectly grouped in phylogeny inference. The great difference in AT usage across Holometabola (Suppl. Table 3, also revealed by RCV and PAUP's Chi-square test) provided strong indication of such compositional heterogeneity across the original matrix. However, the lowered RCV after removal or recoding certain sites confirmed that compositional heterogeneity was mainly confined to the third positions, corroborating this hypothesized parameter as the cause of misleading signal. As an objective measure of the combined effect of saturation and compositional bias we used the quotient of stemminess/RCV, which gives similar results for the two schemes that omitted third positions and eliminated transitions at first codon positions, *i.e.* both $1_{RY}2_{NT}$ and $1_{RY}2_{RY}$ schemes cope similarly well with such confounding effects. Whereas the latter scheme produced a slightly higher stemminess/RCV value, the resulting trees were similar under both schemes, and so we mainly focussed on those obtained with the $1_{RY}2_{NT}$ coding for subsequent analyses, due to the greater number of characters and higher support values.

4.3. Evidence for the Coleoptera–Strepsiptera grouping

While the $1_{RY}2_{NT}$ coding scheme gives the strongest support at several supra-ordinal nodes and recovers Strepsiptera + Coleoptera, the same grouping is repeatedly obtained under alternative coding schemes (Figs. 3–5; Supplement), albeit with lower branch support. Our various phylogenetic analyses recover each insect order as monophyetic with strong support, except for Coleoptera, which in several cases was unresolved with respect to Strepsiptera (Figs. 3–5; Supplement), and *Sphaerius* sp (suborder Myxophaga) most notably difficult to place relative to Strepsiptera. Under the preferred $1_{RY}2_{NT}$ scheme, the possible sister-group relationship of Strepsiptera with Diptera is rejected in statistical tests of topologies (Table 3). Only parsimony analysis of the $1_{RY}2_{RY}$ data weakly placed Strepsiptera as sister-group of Diptera. We suspect this may be due to a failure of parsimony methods to account adequately for heterogeneity, rate shifts and/or multiple substitutions (Carmean and Crespi, 1995; Huelsenbeck, 1997). Model-based analyses of the same $1_{RY}2_{RY}$ data instead again favor a position near Coleoptera, indicating that parsimony analyses might have been misled by long-branch attraction. Each of the three method to phylogeny reconstruction we used have intrinsic weaknesses, but in particular, parsimony is known to perform poorly when a mixture of long-branches affect the topology (*i.e.* long-branch attraction in the "Felsenstein zone"), and where inconsistency leads to the wrong topology with increasing support on addition of more data (Huelsenbeck, 1997, 2001; Hwang et al., 1998). In a similar way, we suspect that long-branch effects may have led to an incorrect placement of Strepsiptera with long-branch Hemiptera with the 2_{NT} data, again, only found here with parsimony methods.

In general, phylogenetic trees from analysis of amino acids were weakly resolved, indicating poor signal at most supra-ordinal nodes. The use of amino acids ameliorates the effect of convergences in nucleotide compositional bias in third codon positions, comparable to R-Y coding and removal of all synonymous codon positions in nucleotide analyses. Yet, as with the 2_{NT} data where Strepsiptera grouped with Hemiptera under parsimony, we suspect the grouping of Strepsiptera and Hemiptera in some protein analyses is also the result of long-branch attraction (Fig. 4 right, Suppl. Fig. 2). In contrast, Strepsiptera again grouped with Coleoptera using mixed models or when site-specific amino acid replacement patterns are accounted for and relative rates were considered a free parameter (CAT-GTR) (Fig. 5 and Suppl. Fig. 3). Analyses of other multi-gene datasets have shown these latter strategies are best able to overcome long-branch attraction effects and avoid systematic errors (Lartillot and Philippe, 2009).

4.4. New insights on the 'Strepsiptera problem' from a revised perspective

Our analyses provide insights to broader relationships across the holometabolan phylogeny. Foremost, these analyses of RP data further validate the Amphiesmenoptera (Trichoptera plus Lepidoptera), perhaps the best-accepted supra-ordinal node of the holometabolan phylogeny (Kristensen, 1997; Whiting, 2004). These analyses also support the placement of Hymenoptera as sister-group to the remaining holometabolan orders, as found in other recent multi-gene analyses (Savard et al., 2006; Wiegmann et al., 2009), and recent morphological analyses of wing characters (reviewed in Beutel and Pohl, 2006; Grimaldi and Engel, 2005).

We generally recovered Amphiesmenoptera together with Diptera and Siphonaptera, *i.e.* the Mecopterida. Within this group, Diptera are usually considered sister-group to Mecoptera + Siphonaptera (=Antliophora), which also is robustly supported by the analysis of six nuclear genes of Wiegmann et al. (2009). In contrast, we found weak support for an alternative sister-group between Diptera and Amphiesmenoptera, to the exclusion of Siphonaptera. We did not recover Antliophora in any of our analyses, which may have been due to the absence of Mecoptera (only Siphonaptera were available), long-branch attraction of Siphonaptera to another lineage, or a combination of such factors. The Antliophora was originally defined by the presence of the male sperm pump, but such structures now appear to be non-homologous in Diptera and Siphonaptera, and absent in some Mecoptera (specifically Nannochoristidae and Boreidae), though monophyly of this latter order is uncertain (Beutel and Pohl, 2006; Grimaldi and Engel, 2005). Despite this, morphological evidence for the Antliophora (without Strepsiptera), comes from recent re-analyses of the male reproductive tract between Diptera, Mecoptera and Siphonaptera, which suggests these three 'mecopteroid' orders share a unique configuration of a U-shaped *vas deferens* that is continuous to the accessory gland, with derived modifications in some Diptera. In contrast, Strepsiptera lack the accessory gland, the *vas deferens* is weakly distinguished from the testes, and seminal vesicles may actually be modified *vasa deferentia* (Hünefeld and Beutel, 2005; Sinclair et al., 2007). Most other putative synapomorphies of Antliophora have now been reinterpreted as reductions or losses that are plesiomorphic for a broader suite of orders (Beutel and Pohl, 2006; Sinclair et al., 2007). Based on our analyses and the recent study of nuclear genes (Wiegmann et al., 2009) a position as sister-group to Coleoptera may now be viewed as the most plausible hypothesis for the Strepsiptera placement (see also Beutel and Gorb, 2001; Kinzelbach and Pohl, 2003; Kukalová-Peck and Lawrence, 1993; Wiegmann et al., 2009). We propose adoption of Coleopterida (Boudreaux, 1979; after Handlirsch, 1903 as Coleopteroidea) for this candidate supra-ordinal group to unite Coleoptera + Strepsiptera (to the exclusion of 'Neuropterida' and remaining orders), while view Elytophora (Packard, 1883) as equivalent to Coleoptera (*sensu* Lankester 1877), wider than the modern usage (*sensu* Latreille 1802).

The placement of Strepsiptera outside Holometabola remains a possibility (see Fig. 4, right), but where such topology is found, we suspect this is due to long-branch attraction. The position near Hemiptera (constraint 14; Table 3) or as sister-group to all Holometabola (constraint 12 and 13); are each clearly rejected with topology comparison tests. A placement of Strepsiptera near Hemiptera is not suggested by morphology, but a few morphological characters support an early branching from all other Holometabola, including the early ontogenetic appearance of compound eyes, the well-developed segment XI in first instar larvae, and the unusual presence of external wing buds in larvae (Beutel and Pohl, 2006; Pohl and Beutel, 2005). One clear limitation of our study was the absence of Neuropterida (=Megaloptera, Raphidioptera, Neuroptera). Evidence for a close affinity between Coleoptera and Neuropterida is provided by shared ovariole morphology (Büning, 2006), but again specialization in Strepsiptera ovarioles prevents clarification of their phylogenetic affinities in this context (Büning, 1998).

If accepted, the close affinity of Strepsiptera with Coleoptera then requires revised interpretation of morphological character state definitions and transformations. An obvious putative synapomorphy for a Coleoptera–Strepsiptera group is opisthomotorism, *i.e.* hindwing motility (Hennig, 1981; Kathirithamby, 1998; Kinzelbach, 1990; Kristensen, 1999; Kukalová-Peck and Lawrence, 1993), elsewhere interpreted as the result of convergent evolution (Wheeler et al., 2001; Whiting, 2002b; Whiting et al., 1997; Whiting and Karthirithamby, 1995). Under the alternative Halteria hypothesis, the haltere-like mesothoracic forewings of Strepsiptera are viewed as homologous to dipteran halteres (*i.e.* metathoracic hindwings), arising from reversed segment identity via a radical homeotic transformation (Whiting and Karthirithamby, 1995; Whiting and Wheeler, 1994). However, such interpretation allows the topological position of a homologous structure to be mutable (Beutel and Pohl, 2006; Hünefeld and Beutel, 2005). In earlier interpretations, strepsipteran forewing modifications were simply deemed convergent with hindwing dipteran halteres, through modifications on different thoracic segments (Nalbach, 1993; Pix et al., 1993). Unlike homeotic transformations and altered segment identity, parallelism or convergence are common evolutionary trends. For example, parallel evolution has been proposed to explain the derived forewings (*i.e.* modified elytra) of certain Coleoptera like *Atractocerus brevicornis* (Lymexylidae), suggested to oscillate during flight much like 'haltere-like' forewings in Strepsiptera (Miller, 1971, see Pix et al., 1993). However, while direct experimentation on the modified forewings of Strepsiptera in a wind tunnel have shown they behave in a functionally similar way to dipteran halteres, the insights on lymexylid forewings are simply speculative (Pix et al., 1993). More commonly, Coleoptera forewings act as passive stabilizers (De Souza and Alexander, 1997). A more defensible example comes from male coccoid Hemiptera (Coccoidea), where hindwings are modified to 'hamulohalteres', anatomically similar to true hindwing halteres of Diptera, but undoubtedly their similarities are due to parallel evolution. More curiously, antennal vibrations in the hawkmoth *Manduca* are also now thought to detect coriolis (gyroscopic) forces during flight (Sane et al., 2007). Plausibly, hawkmoth antennae help maintain stable flight using the same principles as halteres, which detect deviation in the lateral plane (Sane et al., 2007). Consequently, it appears that different morphological structures of insects may be employed to detect deviations in gyroscopic forces, arising through evolutionary convergences. Both mesothoracic or metathoracic wings may adopt a derived role in flight stabilization with relative ease, so that thoracic segments maintain distinct topological identities without invoking radical homeotic transformations.

The Halteria hypothesis has also not stood up to re-evaluation of morphological characters (Carmean and Crespi, 1995; Huelsenbeck, 1997; Kristensen, 1999) or in light of insights from new fossil specimens (Beutel and Pohl, 2006; Pohl et al., 2005). A key facet of the 'Strepsiptera problem' is that few aspects of the derived strepsipteran morphology can unambiguously associate them with any other currently recognized order (Kathirithamby, 1989; Kinzelbach, 1990; Kinzelbach and Pohl, 2003; Beutel and Pohl, 2006). For example, evidence from internal morphology like sperm ultra-structure has proved uninformative about the placement of Strepsiptera (Dallai et al., 2003), while other key aspects of strepsipteran external anatomy are radically different from the condition in other insect orders (Hörnschemeyer, 2002; Kinzelbach and Pohl, 2003; Kristensen, 1999). Most morphological characters previously suggested to link Strepsiptera with Coleoptera involve thoracic modifications, including wing venation and musculature (Kukalová-Peck and Lawrence, 1993), which have since been refuted as mistaken descriptions of strepsipteran wing morphology (Whiting et al., 1997; Whiting and Karthirithamby, 1995). Despite this, support for a Coleoptera–Strepsiptera sister-group can still be found in phylogenetic analyses of morphological data after exclusion of thoracic characters (Beutel and Gorb, 2001; Beutel and Pohl, 2006).

Revised insights on the plesiomorphic strepsipteran morphology have emerged from comparative analyses of fossil specimens preserved in Baltic amber (Pohl et al., 2005), These historical specimens further fuel arguments against an affinity of Strepsiptera to Diptera (Beutel and Pohl, 2006; Krenn, 2007; Pohl et al., 2005). In contrast to extant Strepsiptera, the extinct †*Protoxenos janzeni* has an eight-segmented antenna, a free labrum and robust mandibles (Pohl et al., 2005). The robust mandibles of †*P. janzeni* suggests the dagger-like condition shared between extant Strepsiptera and Antliophora (Whiting et al., 1997) is not a synapomorphy of Halteria (Pohl et al., 2005), but is instead the result of convergence. Consequently, mouthparts provide no unambiguous characters to indicate the phylogenetic position of Strepsiptera (Krenn, 2007). Another potential synapomorphy of Halteria (or Strepsiptera + Antliophora) is the sperm pump (Wheeler et al., 2001; Whiting, 1998; Whiting et al., 1997). Yet here, revisions suggest that insect sperm pumps have multiple independent origins and are unreliable phylogenetic estimators, again weakened by homoplasy (Hünefeld and Beutel, 2005; Sinclair et al., 2007). Consequently, most morphological evidence originally proposed to support the inclusion of Strepsiptera in Antliophora (Whiting et al., 1997; Whiting and Karthirithamby, 1995) has since been critically reinterpreted. Together, revised morphological insights, recent molecular data from other non-rRNA markers, and our analyses presented here, together find little support for Halteria. Instead, the accumulating evidence again indicates the more traditional supra-ordinal Coleoptera + Strepsiptera group as the preferred phylogenetic placement, *i.e.* the Coleopterida.

Acknowledgments

We thank K. Theodorides, A. de Riva and M. Mejia-Chang for generating EST data from Strepsiptera. We are grateful to I. Landais, W. Hunter and J. Eum for providing unpublished EST sequences. We greatly thank Kip Will and three other anonymous reviewers for their comments that greatly improved this manuscript. This work was supported by grants from BBSRC (G14548) and NERC (NER/B/S/2003/00858) to APV, and a BBSRC studentship and subsequent IRCSET fellowship to SJL.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.03.024.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Aspöck, U., 2002. Phylogeny of the Neuropterida (Insecta: Holometabola). Zoologica Scripta 31, 51–55.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., Philippe, H., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proceedings of the National Academy of Sciences USA 99, 1414–1419.

Beutel, R.G., Gorb, S.N., 2001. Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. Journal of Zoological Systematics and Evolutionary Research 39, 177–207.

Beutel, R.G., Pohl, H., 2006. Endopterygote systematics – where do we stand and what is the goal (Hexapoda, Arthropoda)? Systematic Entomology 31, 202–219.

Bonneton, F., Brunet, F.G., Kathirithamby, J., Laudet, V., 2006. The rapid divergence of the ecdysone receptor is a synapomorphy for Mecopterida that clarifies the Strepsiptera problem. Insect Molecular Biology 15, 351–362.

Boudreaux, H.B., 1979. Arthropod Phylogeny with Special Reference to Insects. John Willey & sons, New York, Chichester, Brisbane, Toronto.

Buckley, T.R., Simon, C., Shimodaira, H., Chambers, G.K., 2001. Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. Molecular Biology and Evolution 18, 223–234.

Büning, J., 1998. Reductions and new inventions dominate oogenesis of Strepsiptera (Insecta). International Journal of Insect Morphology and Embryology 27, 3–8.

Büning, J., 2006. Ovariole structure supports sistergroup relationship of Neuropterida and Coleoptera. Arthropod Systematics & Phylogeny 64, 115–226.

Carmean, D., Crespi, B.J., 1995. Do long branches attract flies? Nature Genetics 373, 666.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, 540–552.

Chalwatzis, N., Baur, A., Stetzer, E., Kinzelbach, R., Zimmerman, F.K., 1995. Strongly expanded 18S rRNA genes correlated with a peculiar morphology in the insect order Strepsiptera. Zoology – Analysis of Complex Systems 98, 115–126.

Crowson, R.A., 1960. The phylogeny of Coleoptera. Annual Review of Entomology 5, 111–134.

Dallai, R., Beani, L., Kathirithamby, J., Lupetti, P., Afzelius, B.A., 2003. New findings on sperm ultrastructure of *Xenos vesparum* (Rossi) (Strepsiptera, Insecta). Tissue and Cell 35, 19–27.

De souza, M.M., Alexander, D.E., 1997. Passive aerodynamic stabilization by beetle elytra (wing covers). Physiological Entomology 22, 109–115.

Fiala, K.L., Sokal, R.R., 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution 39, 609–622.

Friedrich, M., Tautz, D., 1997. An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Molecular Biology and Evolution 14, 644–653.

Gillespie, J.J., Mckenna, C.H., Yoder, M.J., Gutell, R.R., Johnston, S., Kathirithamby, J., Cognato, A.I., 2005. Assessing the odd secondary structural properties of nuclear small subunit ribosomal RNA sequences (18S) of the twisted-wing parasites (Insecta: Strepsiptera). Insect Molecular Biology 14, 625–643.

Grimaldi, D., Engel, M.S., 2005. Evolution of the Insects. Cambridge University Press, Cambridge, UK.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52, 696–704.

Handlirsch, A., 1903. Zur Pylogenie der Hexapoden. Akademie der Wissenschaften in Wien, Mathematische-Naturwissenschaftliche Klasse (Abtheilung I) 112, 716–738.

Hayward, D.C., Trueman, J.W.H., Bastiani, M.J., Ball, E.E., 2005. The structure of the USP/RXR of *Xenos pecki* indicates that Strepsiptera are not closely related to Diptera. Development Genes and Evolution 215, 213–219.

Hennig, W., 1953. Kritische bemerkungen zum phylogenetischen system der insekten. Beitrage zur Entomologie 3 (Sonderheft), 1–85.

Hennig, W., 1981. Insect Phylogeny. John Wiley & Sons, Chichester, New York.

Hörnschemeyer, T., 2002. Phylogenetic significance of the wing base of the Holometabola (Insecta). Zoological Scripta 31, 17–29.

Huelsenbeck, J.P., 1997. Is the Felsenstein zone a fly trap? Systematic Biology 46, 69–74.

Huelsenbeck, J.P., 2001. A Bayesian perspective on the Strepsiptera problem. Tijdschrift voor Entomologie 144, 165–178.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Hughes, J., Longhorn, S.J., Papadopoulou, A., Theodorides, K., Riva, A.d., Mejia-Chang, M., Foster, P.G., Vogler, A.P., 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Molecular Biology and Evolution 23, 268–278.

Hünefeld, F., Beutel, R.G., 2005. The sperm pumps of Strepsiptera and Antliophora (Hexapoda). Journal of Zoological Systematics and Evolutionary Research 43, 297–306.

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O.S., Wild, R., Hammond, P.M., Ahrens, D., Balke, M., Caterino, M.S., Gomez-Zurita, J., Ribera, I., Barraclough, T.G., Bocakova, M., Bocak, L., Vogler, A.P., 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science 318, 1913–1916.

Hwang, U.W., Kim, W., Tautz, D., Friedrich, M., 1998. Molecular phylogenetics at the Felsenstein zone: approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. Molecular Phylogenetics and Evolution 9, 470–480.

Kälersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy increases phylogenetic structure. Cladistics 15, 91–93.

Kathirithamby, J., 1989. Review of the order Strepsiptera. Systematic Entomology 14, 41–92.

Kathirithamby, J., 1998. Host-parasitoid associations of Strepsiptera: anatomical and developmental consequences. International Journal of Insect Morphology and Embryology 27, 39–51.

Kinzelbach, R., 1990. The systematic position of Strepsiptera (Insecta). American Entomology 36, 292–303.

Kinzelbach, R.K., Pohl, H., 2003. Ordnung Strepsiptera, Fächerflugler. In: Dathe, H.H. (Ed.), Kaestner – Lehrbuch der Speziellen Zoologie, vol. 1/5: insecta. Spektum Akademischer Verlag, Heidelberg, Berlin, pp. 526–539.

Kjer, K.M., 2004. Aligned 18S and insect phylogeny. Systematic Biology 53, 506–514.

Kjer, K.M., Carle, F.L., Litman, J., Ware, J., 2006. A molecular phylogeny of Hexapoda. Arthropod Systematics and Phylogeny 64, 35–44.

Krauss, V., Thümmler, C., Georgi, F., Lhmann, J., Stadler, P.F., Eisenhardt, C., 2008. Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. Molecular Biology and Evolution 25, 821–830.

Krenn, H.W., 2007. Evidence from mouthpart structure on internodial relationships in Endopterygota? Arthropod Systematics and Phylogeny 65, 7–14.

Kristensen, N.P., 1991. Phylogeny of extant hexapoods. In: RO, C.S.I. (Ed.), The Insects of Australia, second ed. Melbourne University Press, Carlton, Victoria, pp. 125–140.

Kristensen, N.P., 1995. Forty years' insect phylogenetics. Hennig's "Kritische Bemerkungen" and subsequent developments. Zoologische Beitrage NF 36, 83–124.

Kristensen, N.P., 1997. Early evolution of the Trichoptera + Lepidoptera lineage: phylogeny and the ecological scenario. Mémoire Muséum National Histoire Naturelle 173, 253–271.

Kristensen, N.P., 1999. Phylogeny of endopterygote insects, the most successful lineage of living organisms. European Journal of Entomology 96, 237–253.

Kukalová-Peck, J., Lawrence, J.F., 1993. Evolution of the hind wing in Coleoptera. The Canadian Entomologist 125, 181–258.

Landais, I., Ogliastro, M., Mita, K., Nohata, J., López-Ferber, M., Duonor-Cérutti, M., Fournier, P., Devauchelle, G., 2003. Annotation pattern of ESTs from *Spodoptera frugiperda* cells (Sf9) and analysis of the insect-specific features and unexpectedly low codon usage bias. Bioinformatics 19, 2343–2350.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Lartillot, N., Philippe, H., 2009. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. In: Telford, M.J., Littlewood, D.T.J. (Eds.), Animal Evolution, Genomes, Fossils, and Trees. Oxford University Press, pp. 127–138.

Longhorn, S.J., Foster, P.G., Vogler, A.P., 2007. The nematode–arthropod clade revisited: phylogenomic analyses from ribosomal protein genes misled by shared evolutionary biases. Cladistics 23, 130–144.

Marygold, S.J., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Milburn, G.H., Harrison, P.M., Yu, Z., Kenmochi, N., Kaufman, T.C., Leeers, S.J., Cook, K.R., 2007. The ribosomal protein genes and *Minute* loci of *Drosophila melanogaster*. Genome Biology 8, R216.

Miller, P.L., 1971. The possible stabilizing function of the elytra of *Atractocerus brevicornis* (L.) (Lymexylidae: Coleoptera) in flight. The Entomologist 104, 105–110.

Nalbach, G., 1993. The halteres of the blowfly *Calliphora*. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 173, 293–300.

Packard, A.S., 1883. On the classification of the orders of Orthoptera and Neuroptera. The Annals and Magazine of Natural History (ser. 5) 69, 145–154.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular Biology and Evolution 21, 1740–1752.

Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21, 1455–1458.

Phillips, M.J., McLenachan, P.A., Down, C., Gibb, G.C., Penny, D., 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the Major Australasian marsupial radiations. Systematic Biology 55, 122–137.

Pix, W., Nalbach, G., Zel, J., 1993. Strepsipteran forewings are haltere-like organs of equilibrium. Naturwissenschaften 80, 371–374.

Pohl, H., Beutel, R.G., 2005. The phylogeny of Strepsiptera. Cladistics 21, 328–374.

Pohl, H., Beutel, R.G., Kinzelbach, R., 2005. Protoxenidae fam. nov. (Insecta, Strepsiptera) from Baltic amber – a 'missing link' in strepsipteran phylogeny. Zoologica Scripta 34, 57–69.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Rambaut, A., Drummond, A., 2003. Tracer: MCMC Trace Analysis Tool. University of Oxford, Oxford, UK.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., Haeseler, A.v., Kube, M., Reinhardt, R., Burmester, T., 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Molecular Phylogenetics and Evolution 45, 942–951.

Rokas, A., Kathirithamby, J., Holland, P.W.H., 1999. Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. Insect Molecular Biology 8, 527–530.

Sane, S.P., Dieudonne, A., Willis, M.A., Daniel, T.L., 2007. Antennal mechanosensors mediate flight control in moths. Science 315, 863–866.

Savard, J., Tautz, D., Richards, S., Weinstock, G.M., Gibbs, R.A., Werren, J.H., Tettelin, H., Lercher, M.J., 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Research 16, 1334–1338.

Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51, 492–508.

Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16, 1114–1116.

Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17, 1246–1247.

Sinclair, B.J., Borkent, A., Wood, D.M., 2007. The male genital tract and aedeagal components of the Diptera with a discussion of the phylogenetic significance. Zoological Journal of the Linnean Society 150, 711–742.

Sorek, R., Safer, H.M., 2003. A novel method for computational identification of contaminated EST libraries. Nucleic Acids Research 31, 1067–1074.

Swofford, D.L., 2002. PAUP∗: Phylogenetic Analysis using Parsimony. Version 4.0b. Sinauer Associates, Sunderland, MA.

Theodorides, K., Riva, A.d., Gómez-Zurita, J., Foster, P.G., Vogler, A.P., 2002. Comparison of EST libraries from seven beetle species: towards a framework for phylogenomics of the Coleoptera. Insect Molecular Biology 11, 467–475.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25, 4876–4882.

Vogler, A.P., Cardoso, A., Barraclough, T.G., 2005. Exploring rate variation among and within sites in a densely sampled tree: species level phylogenetics of North American tiger beetles (Genus Cicindela). Systematic Biology 54, 4–20.

Wenzel, J.W., Siddall, M.E., 1999. Noise. Cladistics 15, 51–64.

Wiens, J.J., 2003. Incomplete taxa, incomplete characters, and phylogenetic accuracy: what is the missing data problem? Journal of Vertebrate Paleontology 23, 297–310.

Wiegmann, B.M., Trautwein, M.D., Kim, J., Bertone, M., Winterton, S.L., Cassel, B.K., Yeates, D.K., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Evolutionary Biology 7, 34.

Wheeler, W.C., Whiting, M., Wheeler, Q.D., Carpenter, J.M., 2001. The phylogeny of the extant hexapod orders. Cladistics 17, 113–169.

Whitfield, J.B., Kjer, K.J., 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. Annual Review of Entomology 53, 449–472.

Whiting, M.F., 1998. Long branch distraction and the Strepsiptera. Systematic Biology 47, 134–137.

Whiting, M.F., 2002a. Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera. Zoological Scripta 31, 93–104.

Whiting, M.F., 2002b. Phylogeny of the holometabolan insect orders: molecular evidence. Zoological Scripta 31, 3–15.

Whiting, M.F., 2004. Phylogeny of the holometabolous insects. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling the Tree of Life. Oxford University Press, New York.

Whiting, M.F., Carpenter, J.C., Wheeler, Q.D., Wheeler, W.C., 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Systematic Biology 46, 1–68.

Whiting, M.F., Karthirithamby, J., 1995. Strepsiptera do not share hind-wing veinational synapomorphies with the Coleoptera: a reply. Journal of the New York Entomological Society 103, 1–14.

Whiting, M.F., Wheeler, W.C., 1994. Insect homeotic transformation. Nature 368, 696.

Wootton, R.J., 2002. Design, function and evolution in the wings of holometabolan insects. Zoological Scripta 31, 31–40.

Yamanoue, Y., Miya, M., Matsuura, K., Yagishita, N., Mabuchi, K., Sakai, H., Katoh, M., Nishida, M., 2007. Phylogenetic position of tetraodontiform fishes within the higher teleosts: Bayesian inferences based on 44 whole mitochondrial genome sequences. Molecular Phylogenetics and Evolution 45, 89–101.

Zdobnov, E.M., Bork, P., 2007. Quantification of insect genome divergence. Trends in Genetics 23, 16–20.