DOI: 10.1111/j.1365-3113.2006.00355.x

# A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision

STEPHEN L. CAMERON[1], CHRISTINE L. LAMBKIN[2], STEPHEN C. BARKER[3] and MICHAEL F. WHITING[1]

[1]Department of Integrative Biology, Brigham Young University, Provo, Utah, U.S.A.,
[2]Australian National Insect Collection, CSIRO Entomology, Canberra, Australia and
[3]Parasitology Section, School of Molecular and Microbial Sciences, The University of Queensland, Brisbane, Australia

**Abstract.** Mitochondrial genomes provide a promising new tool for understanding deep-level insect phylogenetics, but have yet to be evaluated for their ability to resolve intraordinal relationships. We tested the utility of mitochondrial genome data for the resolution of relationships within Diptera, the insect order for which the most data are available. We sequenced an additional three genomes, from a syrphid, nemestrinid and tabanid, representing three additional dipteran clades, 'aschiza', non-heteroneuran muscomorpha and 'basal brachyceran', respectively. We assessed the influence of optimality criteria, gene inclusion/exclusion, data recoding and partitioning strategies on topology and nodal support within Diptera. Our consensus phylogeny of Diptera was largely consistent with previous phylogenetic hypotheses of the order, except that we did not recover a monophyletic Muscomorpha (Nesmestrinidae grouped with Tabanidae) or Acalyptratae (Drosophilidae grouped with Calliphoridae). The results were very robust to optimality criteria, as parsimony, likelihood and Bayesian approaches yielded very similar topologies, although nodal support varied. The addition of ribosomal and transfer RNA genes to the protein coding genes traditionally used in mitochondrial genome phylogenies improved the resolution and support, contrary to previous suggestions that these genes would evolve too quickly or prove too difficult to align to provide phylogenetic signal at deep nodes. Strategies to recode data, aimed at reducing homoplasy, resulted in a decrease in tree resolution and branch support. Bayesian analyses were highly sensitive to partitioning strategy: biologically realistic partitions into codon groups produced the best results. The implications of this study for dipteran systematics and the effective approaches to using mitochondrial genome data are discussed. Mitochondrial genomes resolve intraordinal relationships within Diptera accurately over very wide time ranges (1–200 million years ago) and genetic distances, suggesting that this may be an excellent data source for deep-level studies within other, less studied, insect orders.

Correspondence: Stephen L. Cameron, Department of Integrative Biology, Brigham Young University, Provo, UT 84602, U.S.A. E-mail: stephen_cameron@byu.edu

## Introduction

Considerable effort is currently being expended to develop new classes of molecular data for use in insect systematics, with the view that many of the higher level relationships (inter- and intraordinal level) cannot be resolved satisfactorily with the current limited array of molecular markers which have been optimized for routine sequencing. Of the vast range of potential genes which could be used for phylogenetic inference, only the nuclear ribosomal RNA (rRNA) genes (*18S* and *28S*), some of the histone subunits (principally *H3*), portions of the mitochondrial (mt) genes (*cox1*, *cox2*, *cytB*, *16S* and *12S*) and a few developmental genes, such as wingless (*Wg*) and Hox (*Hx*), are sufficiently conservative to be sequenced readily across all orders (Caterino *et al.*, 2000). Other genes, such as opsin, elongation factors and rudimentary (CAD), have been optimized for some orders, but have proven difficult to use across multiple orders (Moulton & Wiegmann, 2004). New markers are being pursued to expand this array of tools to improve phylogenetic resolution between and within orders.

One potential marker is the mt genome. With thirty-seven genes [thirteen protein coding, two rRNAs and twenty-two transfer RNAs (tRNAs)], and usually between 14 000 and 17 000 base pairs in size, the mt genome of Metazoa is the smallest extant genome. This size makes it technically tractable to sequence in its entirety, but it is still an order of magnitude larger than most of the single genes used in current insect phylogenetic analyses (Saccone *et al.*, 1999). Two approaches to the use of the mt genome have been proposed: uncovering shared genome rearrangements and use of the whole genome in sequence-based phylogenies (e.g. Boore & Brown, 1998; Rokas & Holland, 2000). Because some of the first insect mt genomes to be sequenced, i.e. *Apis mellifera*, Hymenoptera (Crozier & Crozier, 1993), *Locusta migratoria*, Orthoptera (Flook *et al.*, 1995a) and *Heterodoxus macropus*, Phthiraptera (Shao *et al.*, 2001) had gene order rearrangements relative to the inferred ancestral mt gene order, early efforts focused on the use of 'genome morphology' (Dowton *et al.*, 2002) to resolve deep nodes in insect evolution. Subsequent sequencing effort has not supported this notion. Although genome rearrangements have shown signal within some orders, e.g. Orthoptera (Flook *et al.*, 1995b), Hymenoptera (Dowton & Austin, 1999; Dowton *et al.*, 2003) and Phthiraptera (Covacin *et al.*, 2006), there are insufficient rearrangements in most genomes to resolve phylogenies adequately. Simultaneous with this realization has been a shift in emphasis to using the mt genome sequence in phylogenetic reconstruction.

The use of whole mt genome sequences in insect phylogenetics has ranged from resolving strains within the *Drosophila melanogaster* subgroup (Ballard, 2000a) to phylogenies of all arthropods (e.g. Nardi *et al.*, 2003a; Cameron *et al.*, 2004). These phylogenies have produced some remarkable results, such as inferring Phthiraptera + Hymenoptera (Nardi *et al.*, 2003a), the polyphyly of Hexapoda (Nardi *et al.*, 2003a; Bae *et al.*, 2004) and Orthoptera + Holometabola to the exclusion of Paraneoptera (Stewart & Beckenbach, 2003). These results have highlighted the need for rigorous evaluation of the phylogenetic behaviour of mt genomes to increase confidence that results obtained from this marker reflect evolutionary history. Cameron *et al.* (2004) evaluated the effects of outgroup choice, data transformations (DNA to amino acids), restricted gene analyses and optimality criteria on the behaviour of mt genome phylogenies of the basal hexapod relationships, in particular the relationships of Collembola to Insecta. This study demonstrated that, although the relationships of the collembolans were very sensitive to variations in all of these factors, there was remarkable consistency in the intraordinal relationships of the representatives of Diptera, Lepidoptera and Coleoptera. This hinted that the mt genome might be particularly valuable in resolving phylogenetic relationships within insect orders. We test this proposition with the insect order for which the most mt genomes have been sequenced, Diptera.

Diptera is the perfect group in which to test the utility of mt genomes in insect intraordinal phylogenetic reconstruction for several reasons. First, many genomes have already been sequenced, with eleven species in four families providing one of the largest representations of any invertebrate group in GenBank. Furthermore, these representatives are spread over three of the major groups within the order: 'nematocera',* 'acalyptrata' and Schizophora. The only missing groups represent grades that lie between the 'nematocerans' and the schizophorans: the 'basal brachycerans' and 'aschizans'. Thus sequencing only a few additional families provides a relatively complete representation of significant groups. By contrast, for most insect orders, only a single mt genome has been sequenced, and, for those orders in which multiple representatives are available, frequently these are very close relatives. For example, Lepidoptera, one of the better studied orders, is represented by five species from three genera in two families, but both families are members of the highly derived clade, Obtectomera (Kristensen & Skalski, 1999). Secondly, the deeper level phylogeny of Diptera has received considerable attention in recent years to provide a large background of phylogenetic data against which mt genome-derived trees can be compared. This background phylogenetic data from a variety of sources, including comprehensive morphological and molecular analyses utilizing multiple nuclear genes, are highly congruent between the different data sources (Yeates & Wiegmann, 1999, 2005). A range of deep-level splits in the dipteran tree are well supported as monophyletic (e.g. Brachycera, Cyclorrhapha, Schizophora) and can be used for comparisons of the utility of mt genome data. A better understanding of the utility and limits of mt genome data will, in turn, allow the confident application of mt genome data to groups such as Coleoptera or Lepidoptera, in which there is much less consensus about phylogenetic relationships (Kristensen & Skalski, 1999; Vogler, 2005).

Here, we present the results of phylogenetic analyses using the Diptera mt genome. We have sequenced representatives

*Many taxonomic groups recognized previously within Diptera are now known to be paraphyletic evolutionary grades. Throughout this paper, these terms have been retained to allow ready comparison with older literature and are designated by being placed in quotation marks.

of three additional families, Tabanidae, Nemestrinidae and Syrphidae, which belong to three grades of dipteran evolution, 'basal brachycera', non-heteroneuran muscomorphs and 'aschiza', respectively. Complete mt genomes of these taxa, previously unrepresented in GenBank, are pivotal to the evaluation of the ability of mt genomes to resolve dipteran phylogeny accurately because they bridge the phylogenetic gap between early branching 'nematocerans' and the more derived schizophorans. We evaluate the influence of data transformations and optimality criteria on the recovered topologies and nodal support, and estimate the timescales over which mt genome data are informative, to gain a better understanding of the utility of mt genome sequence phylogenies and the types of questions to which they may be confidently applied.

## Materials and methods

### *Mitochondrial genome sequencing*

Specimens of the tabanid, *Cydistomyia duplonotata* (Ricardo), and nemestrinid, *Trichophthalma (Lichtwardtiomyia) punctata* (Macquart), were collected from Benarkin State Forest (near Blackbutt, south-east Queensland, Australia) by S. Cameron, J. Nielsen, M. Rix and G. Svenson, on 8 December 2001, and of the syrphid, *Simosyrphus grandicornis* (Macquart), from Cedar Creek (Pine Rivers Shire, south-east Queensland, Australia) by S. Cameron, on 3 October 2001. All specimens were snap frozen in liquid nitrogen and stored at −80° C in the insect tissue collection of the Department of Integrative Biology, Brigham Young University. Flies were identified by C. Lambkin and voucher specimens have been deposited in the Australian National Insect Collection, accession numbers: ANIC29-009567 (*T. punctata*), ANIC29-009568 (*S. grandicornis*) and ANIC29-009569 (*C. duplonotata*).

Whole genomic DNA was extracted from thoracic muscle tissue with the DNeasy Tissue Kit (Qiagen, Valencia, CA). Short regions of the *cox1*, *cox2*, *cytB*, *12S* and *16S* genes were amplified using general insect primers (Table S1, 'Supplementary material') and sequenced. These short sequenced regions were used to design specific primers which, in combination with general insect primers, allowed the amplification of the whole genome of each species by long polymerase chain reaction (PCR). The primer sequence and location for each long PCR are listed in Table S2 ('Supplementary material'). Within each long PCR product, the full double-stranded sequence was determined by primer walking (primers available from S. Cameron on request). Short PCRs were performed using Elongase (Invitrogen, Carlsbad, CA) with the following cycling conditions: 95 °C for 12 min; forty cycles of 94 °C for 1 min, 40 °C for 1 min and 72 °C for 1 min; and a final elongation of 72 °C for 7 min. Long PCRs were performed using Elongase with the following cycling conditions: 92 °C for 2 min; forty cycles of 92 °C for 30 s, 50 °C for 30 s and 68 °C for 12 min; and a final run out step of 68 °C for 20 min. Sequencing was performed using ABI (Foster City, CA) BigDye version 3 dye terminator sequencing technology and run on an ABI 3770 or ABI 3740 capillary sequencer. Sequencing PCR conditions were twenty-eight cycles of 94 °C for 10 s, 50 °C for 5 s and 60 °C for 4 min. An ambiguous section of the *T. punctata* mt genome between *nadh4* and *nadh6* was resolved by cloning this region using Topo-TA cloning chemistry (Invitrogen).

Raw sequence files were edited and assembled into contigs in Sequencher version 4 (GeneCodes Corporation, Ann Arbor, MI). tRNA analysis was conducted using tRNAscan-SE (Lowe & Eddy, 1997) employing invertebrate mt predictors and a COVE score cut-off of unity. Reading frames between tRNAs were found in Sequencher and identified using translated BLAST searches (BLASTX) (Altschul *et al.*, 1997) as implemented by the NCBI website (http://www.ncbi.nlm.nih.gov/).

### *Testing regime*

To test the limits and utility of the mt genome in accurate intraordinal reconstructions, we assessed variation in three areas. First, the influence of inference methodology was tested by performing replicate analyses under maximum parsimony (MP), maximum likelihood (ML) and Bayesian (BA) analyses. Consistency across methods was assessed by comparison of topological differences between the methods and of the strength of nodal support for those nodes in common. Secondly, effects of data partitioning were examined. Comparisons of topology and nodal support were performed within each of the three inference frameworks to test consistency between the three main data partitions – protein coding genes (PCG), rRNA genes (RIBO) and tRNA genes (TRAN) – and also in combined analyses which included all three partitions (ALL). The possible effect of substitution saturation at the third codon position of the protein coding genes was assessed by comparison of datasets including (DNA123) or excluding (DNA12) third codon positions, and of those in which the PCG partition had been recoded as amino acids (PROT); these comparisons were performed for both the PCG partition alone and the combined ALL partition. The optimal partitioning strategy for BA analysis was assessed by comparison of Bayes factors (Nylander *et al.*, 2004) for runs partitioned by gene, by codon and by codon and gene. Gene 'quality' was assessed by comparison of homoplasy levels across gene partitions in MP analyses following our previous methods (Cameron *et al.*, 2004; see below). Thirdly, the time range over which mt genomes provided reliable phylogenetic signal was assessed by comparison of maximum nodal age from fossil evidence and previous dating analyses. By assessing which nodes were stable and well supported across all inference and partition methods, we were able to calculate upper and lower dates for which we were confident that phylogenetic reconstructions were accurate. A synopsis of the analyses performed is included in Table 1.

**Table 1.** Analyses performed in this study.

| Dataset | Subanalyses | Parsimony (MP) | Likelihood (ML) | Bayesian (BA) |
|---|---|---|---|---|
| ALL-DNA123 | | Yes | Yes | |
| | Gene partitions (GP) | | | Yes |
| | Codon partitions (CP) | | | Yes |
| | Codon & gene partitions (CGP) | | | Yes |
| ALL-DNA12 | | Yes | Yes | |
| | Gene partitions (GP) | | | Yes |
| | Codon partitions (CP) | | | Yes |
| | Codon & gene partitions (CGP) | | | Yes |
| ALL-PROT | | Yes | | Yes |
| PCG-DNA123 | | Yes | Yes | |
| | Gene partitions (GP) | | | Yes |
| | Codon partitions (CP) | | | Yes |
| | Codon & gene partitions (CGP) | | | Yes |
| PCG-DNA12 | | Yes | Yes | |
| | Gene partitions (GP) | | | Yes |
| | Codon partitions (CP) | | | Yes |
| | Codon & gene partitions (CGP) | | | Yes |
| PCG-PROT | | Yes | | Yes |
| RIBO | | Yes | Yes | Yes |
| TRAN | | Yes | Yes | Yes |

ALL, all gene partitions; DNA123, nucleotide sequences, all codon positions; DNA12, nucleotide sequences, first and second positions only; PCG, protein coding genes only; PROT, amino acid sequences; RIBO, ribosomal RNA genes only; TRAN, transfer RNA genes only.

### *Phylogenetic inference*

All the currently available mt genomes of Holometabola were used (see Table 2),† including multiple representatives of the Lepidoptera (five species in three families), Coleoptera (three species in three families) and Hymenoptera (three species in two families). *Tricholepidion* (Zygentoma) and *Triatoma* (Hemiptera) were used as outgroups.

An amino acid alignment was generated in ClustalW (Thompson *et al.*, 1994; implemented in mega3; Kumar *et al.*, 2004) for each of the thirteen protein coding genes, and a DNA alignment was inferred from the amino acid alignment using mega3 (Kumar *et al.*, 2004), which can translate between DNA and amino acid sequences within alignments. Amino acid alignments were made with the following parameters: pairwise alignment gap opening penalty = 10 and extension penalty = 0.1; multiple alignment gap opening penalty = 10 and extension penalty = 0.2; protein weight matrix = Gonnet; residue specific penalties: on; hydrophilic penalties: on; gap separation distance = 4; end gap separation: off; negative matrix = off; and delay divergent cut-off = 30%. tRNA and rRNA genes were aligned in mega3 using ClustalW, and corrected by eye to account for their secondary structures using the following parameters: gap opening cost = 15; gap extension cost = 6.66; DNA weight matrix = IUB; transition weight cost = 0.5; negative matrix = off; and delay divergent cut-off = 30%. Alignments of individual genes were then concatenated in MacClade 4.06 (Maddison & Maddison, 2003), and data partitions were delimited on the basis of each included gene and for each codon position. Each of the rRNA genes was included as a separate partition, but the tRNA genes were joined in a single partition for ModelTest and BA, as the number of variable sites within individual tRNAs was too few for accurate parameter calculations.

Phylogenetic analysis was performed with paup 4.0b10 (Swofford, 2002) for MP and ML analyses and with MrBayes (BA) versions 3.0b4 and 3.1.1 (Huelsenbeck & Ronquist, 2001) for BA analysis. Bootstrap support was calculated with paup 4.0b10 with either 1000 (MP) or 100 (ML) replicates. Tree statistics were calculated in paup 4.0b10. The relative contribution of each dataset partition to the combined topology was calculated using partitioned Bremer (Baker & DeSalle, 1997), as implemented in TreeRot version 2 (Sorenson, 1999). All BA analyses were run with unlinked partitions, appropriate models of molecular evolution selected for each partition and analyses run using four chains, for three million generations with sampling every 1000 generations. Completed BA analyses were examined for asymptotic behaviour of each parameter and for total tree likelihood; trees collected prior to this asymptotic point were treated as burn-in and discarded (generally the first 30 000–60 000 generations). Models for BA and ML were chosen using AIC as implemented in ModelTest (Posada & Crandall, 1998), run independently for each partition used in BA and for combined partitions in ML analyses. ML and BA run files are available for all analyses on request.

†Three additional dipteran mitochondrial genomes have become available recently on GenBank – *Dermatobia hominis* (Oestridae), *Haematobia irritans* (Muscidae) and *Aedes albopictus* (Culicidae) – but have not yet been published. We have not included these sequences so as not to pre-empt other workers. Further, as close relatives of each have already been published, their inclusion is not critical to this study.

**Table 2.** Taxon sampling and availability.

| Species | Order | Family | Accession number | Availability |
|---|---|---|---|---|
| *Tricholepidion gertischi* | Zygentoma | Tricholepidididae | NC005437 | Nardi *et al.* (2003a) |
| *Triatoma dimidiata* | Hemiptera | Reduvidae | NC002609 | Dotson & Beard (2001) |
| *Tribolium castaneum* | Coleoptera | Tenebrionidae | NC003081 | Friedrich & Muqim (2003) |
| *Crioceris duodecimpuctata* | Coleoptera | Chrysomelidae | NC003372 | Stewart & Beckenbach (2003) |
| *Pyrocoelia rufa* | Coleoptera | Lampyridae | NC003970 | Bae *et al.* (2004) |
| *Antheraea pernyi* | Lepidoptera | Saturnidae | NC004622 | Liu *et al.*, unpublished |
| *Bombyx mori* | Lepidoptera | Bombycidae | NC002355 | Lee *et al.*, unpublished |
| *Bombyx mandarina* | Lepidoptera | Bombycidae | NC003395 | Yukuhiro *et al.* (2002) |
| *Ostrinia furnacalis* | Lepidoptera | Crambidae | NC003368 | Coates *et al.* (2005) |
| *Ostrinia nubialis* | Lepidoptera | Crambidae | NC003367 | Coates *et al.* 2005 |
| *Anopheles quadrimaculatus* | Diptera | Culicidae | NC000875 | Mitchell *et al.* (1993) |
| *Anopheles gambiae* | Diptera | Culicidae | NC002084 | Beard *et al.* (1993) |
| *Cydistomyia duplonotata* | Diptera | Tabanidae | DQ866052 | Present study |
| *Trichophthalma punctata* | Diptera | Nemestrinidae | DQ866051 | Present study |
| *Simosyrphus grandicornis* | Diptera | Syrphidae | DQ866050 | Present study |
| *Drosophila yakuba* | Diptera | Drosophilidae | NC001322 | Clary & Woolstenholme (1985) |
| *Drosophila melanogaster* | Diptera | Drosophilidae | NC001709 | Lewis *et al.* (1995) |
| *Drosophila simulans* | Diptera | Drosophilidae | NC005781 | Ballard (2000a) |
| *Drosophila sechellia* | Diptera | Drosophilidae | NC005780 | Ballard (2000a) |
| *Drosophila mauritiania* | Diptera | Drosophilidae | NC005719 | Ballard (2000b) |
| *Chrysomya putoria* | Diptera | Calliphoridae | NC002697 | Junqueira *et al.* (2004) |
| *Cochliomyia hominovorax* | Diptera | Calliphoridae | NC002660 | Lessinger *et al.* (2000) |
| *Ceratitis capitata* | Diptera | Tephritidae | NC000857 | Spanos *et al.* (2000) |
| *Bactrocera oleae* | Diptera | Tephritidae | NC005333 | Nardi *et al.* (2003b) |
| *Perga condei* | Hymenoptera | Pergidae | AY787816 | Castro & Dowton (2005) |
| *Melipona bicolor* | Hymenoptera | Apidae | NC004529 | Silvestre & Arias, unpublished |
| *Apis mellifera* | Hymenoptera | Apidae | NC001566 | Crozier & Crozier (1993) |

The distribution of homoplasy amongst the genes was estimated by gene tree/total tree comparisons (Cameron *et al.*, 2004). The proportional size of each gene (length of the aligned gene partition divided by the total alignment length) was compared with its proportional contribution to tree length (tree length for that gene partition determined by TreeRot divided by the total tree length for the combined analysis). Thus, the formulae is $\mathrm{nl_g/nl_t : tl_g/tl_t}$, where $\mathrm{nl_g}$ is the aligned length of a particular gene, $\mathrm{nl_t}$ is the nucleotide length of the total alignment, $\mathrm{tl_g}$ is the partitioned tree length for that gene and $\mathrm{tl_t}$ is the total tree length. Replicate analyses were performed using total gene length or informative sites alone. If homoplasy is distributed randomly within the dataset, these proportions should be the same, whereas, if homoplasy is clustered in particular genes, these should contribute extra tree length to the combined analysis disproportionate to the length of the gene. In this way, the relative contributions of different data partitions can be compared directly.

## Results

### Genome sequences

The entire mt genomes of *Cydistomyia duplonotata* (tabanid), *Trichophthalma (Lichtwardtiomyia) punctata* (nemestrinid) and *Simosyrphus grandicornis* (syrphid) were sequenced by primer walking of multiple overlapping long PCR fragments. These genomes have been deposited in GenBank with the following accession numbers: *C. duplonotata* DQ866052, *T. punctata* DQ866051 and *S. grandicornis* DQ866050. The genomes are 16 247, 16 396 and 16 141 base pairs long, respectively. Each genome has the usual metazoan complement of thirteen protein coding genes, two rRNA genes and twenty-two tRNA genes. Neither of the genome rearrangements previously found in Diptera, tRNAs RANSEF found in Culicidae (Beard *et al.*, 1993; Mitchell *et al.*, 1993) or duplication of tRNAs I and Q found in Calliphoridae (Lessinger *et al.*, 2004) were present. Each of the three genomes has the plesiomorphic pancrustacean mt genome arrangement (Boore, 1999). cove analysis failed to find the Arg tRNA for any of the three species and the Ser (UCN) tRNA in *Cydistomyia*; these genes were identified by aligning the region expected to be occupied by these tRNAs against homologous regions from other insects. In addition, cove analysis misidentified the tRNA isoform of Ser (UCN) as Phe as a result of misplacement of the anticodon loop; this was identified properly by alignments of this tRNA amongst Diptera and other insects. Additional tRNAs were detected in the AT-rich regions (= putative control region, origin of replication) of each species. However, as there appeared to be no evolutionary conservation of these regions across the taxa and the cove scores were very low, it is probable that these are spurious tRNAs: regions of random sequence capable of folding into clover-leaf structures similar to real tRNAs. Despite fairly large AT-rich regions of 1376, 1597 and 1127 bp, respectively, large repetitive sequences were absent. Each of the thirteen protein coding genes had the regular start (I or M) and stop (TAG, TAA, TA or T) codons, except for *cox1* in the tabanid and nemestrinid, which did not have an available start codon. The start codons were identified by comparison of alignments with other dipterans to be R and Q, respectively. Start codons for *cox1* are frequently non-standard in holometabolans, and so this finding is not particularly unusual (Lessinger *et al.*, 2000; Junqueira *et al.*, 2004). Drosophilids utilize tetranucleotide start codons for *cox1* (ATAA = S) and other dipteran groups appear to simply use the first in-frame codon which is usually an R, S, T or Q.

### *Combined Dipteran mitochondrial genome phylogeny*

We performed thirty separate phylogenetic analyses to test the effects of the optimality criterion, data partitioning strategies and data coding approaches on mt genome phylogenies of Diptera. Tree statistics are presented in Table 3. The results of analysis of the ALL-DNA12 dataset under the three optimality criteria are presented in Figs 1–3. As the majority of analyses recovered results which were largely congruent with the trees recovered using the maximal amount of genome sequence data, the results of other analyses are not presented. Tables S3–S6 ('Supplementary material') depict which nodes are congruent and the level of nodal support. A full complement of tree data (alignments plus tree figures) is available from the Whiting laboratory website (http://whitinglab.byu.edu/). There is remarkable congruence between datasets and across optimality criteria, with most of the intraordinal nodes recovered from a majority of analyses. Holometabola is monophyletic in all except some PCG datasets. Interordinal relationships vary widely, with all possible combinations of the four holometabolous orders recovered in at least some analyses. The orders generally are monophyletic, except in the RIBO analyses, in which Coleoptera becomes a paraphyletic grade at the base of Holometabola and *Melipona* generally groups with the tephritids, rendering most dipteran clades non-monophyletic. Diptera is monophyletic in all except two analyses: RIBO-ML and RIBO-BA. These relationships found in the RIBO datasets are probably a result of alignment artefacts caused by the large indels in these genes. Intraordinal relationships within Diptera are also broadly congruent across analyses. The major dipteran clades, Brachycera, Cyclorrhapha and Schizophora, are recovered in almost all analyses, and intrafamilial and congeneric relationships are universally recovered. Furthermore, the overall topology is broadly congruent with current consensus phylogenies of the Diptera (Wiegmann *et al.*, 2003; Yeates & Wiegmann, 2005). The topology differs in two areas, the 'basal brachycera' and the 'acalyptrates'. Our trees recover monophyly for the 'basal brachycerans', whereas the Tabanidae and Nemestrinidae are placed more usually in infraorders Tabanomorpha and Muscomorpha, respectively, with the nemestrinids as the earliest branch within Muscomorpha (Wiegmann *et al.*, 2000, 2003; Yeates, 2002). By contrast, 'acalyptrate' monophyly is controversial, asserted by many authorities but lacking morphological support (Yeates & Wiegmann, 2005).

### *Optimality criteria*

The major effect of the optimality criteria is seen at the interordinal level. MP divides Holometabola into two groups: Coleoptera and an unresolved trichotomy of Lepidoptera, equivalent for the included taxa to support for a monophyletic Antliophora. ML splits Holometabola into two groups: Coleoptera + Hymenoptera and Lepidoptera + Diptera. BA splits Diptera from the remaining Holometabola. It should be noted, however, that interordinal relationships vary much more between datasets than between criteria, and dataset is probably the more significant determining variable (see below). By contrast, optimality criteria have almost no influence within orders, and particularly the intraordinal relationships within Diptera, which are almost

**Table 3.** Tree statistics.

| | Parsimony | | | | | Likelihood | Bayesian[a] |
|---|---|---|---|---|---|---|---|
| Dataset | No. sites | Length | CI | RI | RC | −ln L | Log. likelihood |
| ALL-DNA123 | 15803 | 46461 | 0.42 | 0.49 | 0.21 | 197367.4 | 189742.8 |
| ALL-DNA12 | 11888 | 26854 | 0.48 | 0.56 | 0.27 | 122448.3 | 119137.0 |
| ALL-PROT | 7956 | 24338 | 0.58 | 0.58 | 0.34 | n/a | 45706.6 |
| PCG-DNA123 | 11745 | 36879 | 0.41 | 0.48 | 0.19 | 154836.1 | 147642.3 |
| PCG-DNA12 | 7830 | 17270 | 0.48 | 0.56 | 0.27 | 79920.4 | 77579.7 |
| PCG-PROT | 3898 | 14708 | 0.64 | 0.62 | 0.39 | n/a | 84781.4 |
| RIBO | 2391 | 6338 | 0.48 | 0.54 | 0.26 | 25736.7 | 25814.2 |
| TRAN | 1667 | 3161 | 0.51 | 0.58 | 0.30 | 14941.6 | 14992.0 |

ALL, all gene partitions; DNA123, nucleotide sequences, all codon positions; DNA12, nucleotide sequences, first and second positions only; PCG, protein coding genes only; PROT, amino acid sequences; RIBO, ribosomal RNA genes only; TRAN, transfer RNA genes only; CI, consistency index; RI, retention index; RC, rescaled consistency index.

[a]For the ALL and PCG datasets, the partition by codon and gene datasets is given as these are the most likely.

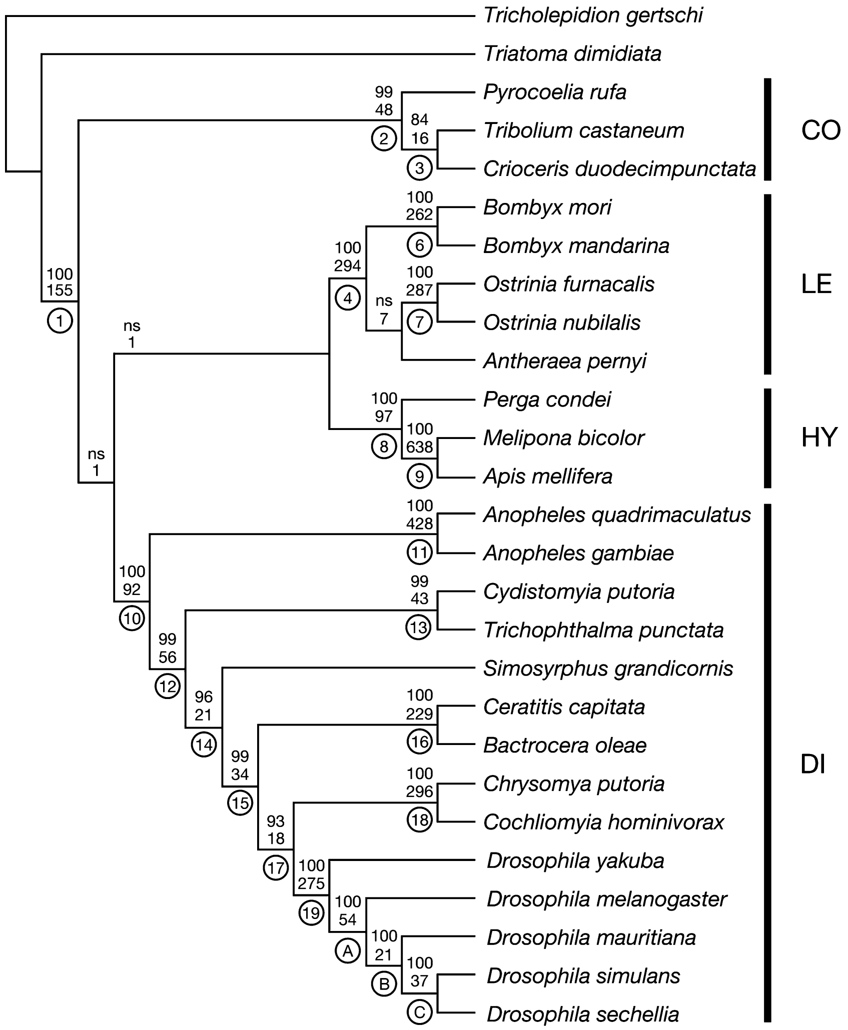

**Fig. 1.** Parsimony cladogram inferred from the ALL-DNA12 dataset. Support is indicated above each node: the top number is the bootstrap percentage, the bottom number the non-partitioned Bremer support. Numbers below each node (circled) are common across datasets; partitioned Bremer supports for each gene are given in Table S3; support levels across different datasets are given in Table 7. CO, Coleoptera; DI, Diptera; HY, Hymenoptera; LE, Lepidoptera; ns, not supported.

constant across the optimality frameworks for comparable datasets (schizophoran paraphyly in the ML-DNA123 dataset is the only exception). With regard to nodal support, across datasets there was a predictable increase in support levels comparing results from MP to ML to BA. As previously noted by many workers (e.g. Leaché & Reeder, 2002; Whittingham *et al.*, 2002; Cameron *et al.*, 2004), BA posterior probabilities are inflated relative to both ML and MP bootstraps. Indeed, several datasets recovered topologies for which every node had a posterior probability of 1.0. Differences between the ML and MP bootstrapping can be related to the greater sensitivity of the latter to homoplasy. mt datasets, particularly nucleotide analyses which include the third codon position of protein coding genes, have fairly high levels of homoplasy. Given the common topology within Diptera, the differences in nodal support between the three optimality criteria probably can be interpreted as upper (BA), middle (ML) and lower (MP) confidence bounds to the inferred relationships.

*Effects of data partitioning*

Data partitioning strategies had much larger effects than optimality criteria on both topology and support. Compared with the ALL analyses, the three datasets utilizing reduced numbers of genes (PCG, RIBO, TRAN) were less likely to recover the same topology and, for those nodes which were recovered in common, nodal support was usually reduced. The ALL datasets were a third (DNA123) to twice (PROT) as large as the PCG datasets (Table 3), and the magnitudes of the differences between the ALL and PCG analyses were related to this relative increase in data, i.e. ALL-PROT differed from PCG-PROT more than ALL-DNA123 did from PCG-DNA123 in both

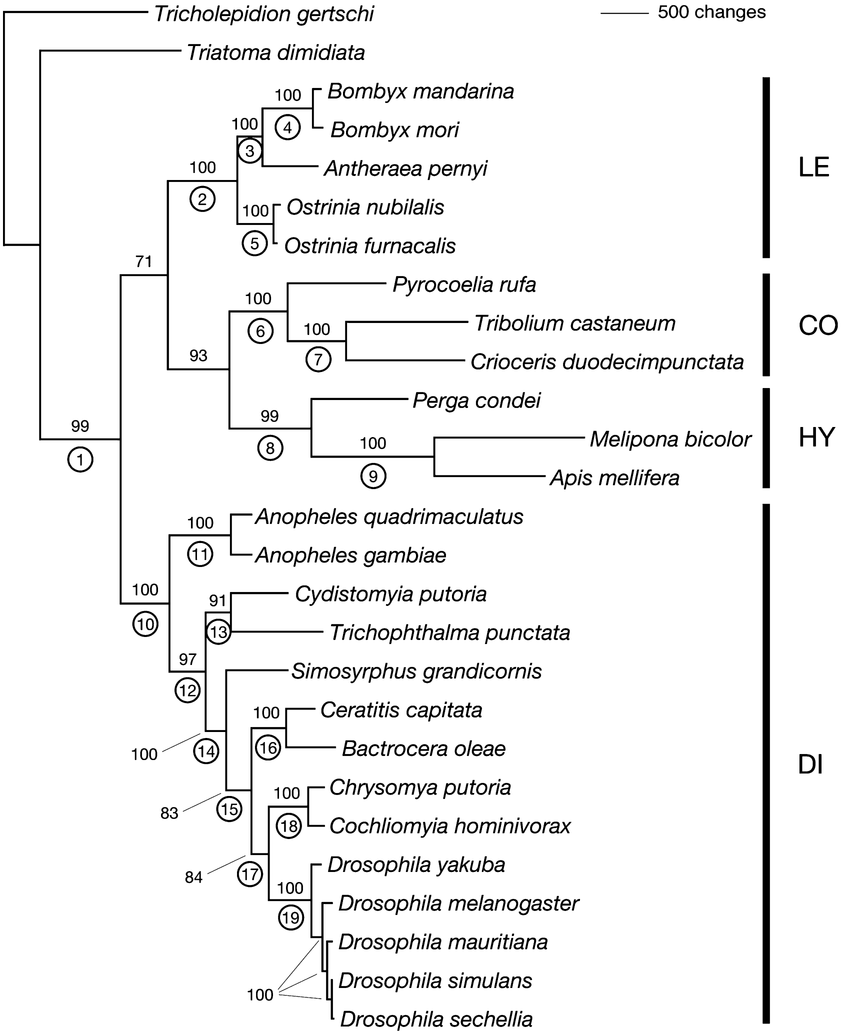

**Fig. 2.** Likelihood phenogram inferred from the ALL-DNA12 dataset. Bootstrap percentages are indicated above each node. Numbers below each node (circled) are common across datasets; support levels across different datasets are given in Table S4. CO, Coleoptera; DI, Diptera; HY, Hymenoptera; LE, Lepidoptera.

topology and nodal support (Tables S3–S6; 'Supplementary material'). In line with the dogma of total evidence (Kluge, 1989), more data appear to yield more robust phylogenetic estimates. The surprising exception is in the comparison of the RIBO and TRAN datasets. The RIBO dataset is almost 50% larger than the TRAN dataset (2391 bp vs. 1667 bp), and yet the TRAN dataset consistently outperforms RIBO. The RIBO topologies are the worst of the analyses performed here. Coleoptera is a paraphyletic grade at the base of Holometabola. *Pyrocoelia* and *Melipona* have wildly incorrect positions in the tree branching within Lepidoptera and Diptera, respectively, and both positions actually receive reasonable nodal support. These problematic topologies may be a function of the unsophisticated alignment procedure used (Clustal), which is known to struggle with ribosomal sequence (e.g. Terry & Whiting, 2005), or long-branchlike effects caused by base composition and rate heterogeneities within insect mt rRNA genes, or both. By contrast, the TRAN datasets, with less data, produced much more reasonable phylogenies, particularly for deep nodes. The orders were monophyletic and, with the exception of Hymenoptera, received significant nodal support in all analyses. The highly conserved structural qualities of the tRNAs, plus consistent annotational methodology by mt genomics researchers, ensuring that truly homologous regions are compared, have probably contributed to the phylogenetic stability of these genes. Conflict of the TRAN datasets lies within Diptera where the topology differs considerably from both the ALL datasets and previous expectation. The position of the syrphid is variable, either as first branch within Diptera (MP) or as sister to *Anopheles* (ML, BA). The positions of the tabanid and nemestrinid are also variable, a paraphyletic grade at the base of Schizophora (MP) or a monophyletic sister group to *Anopheles* + syrphid. Relationships between the three Schizophoran families also vary, with Tephritidae + Calliphoridae rather

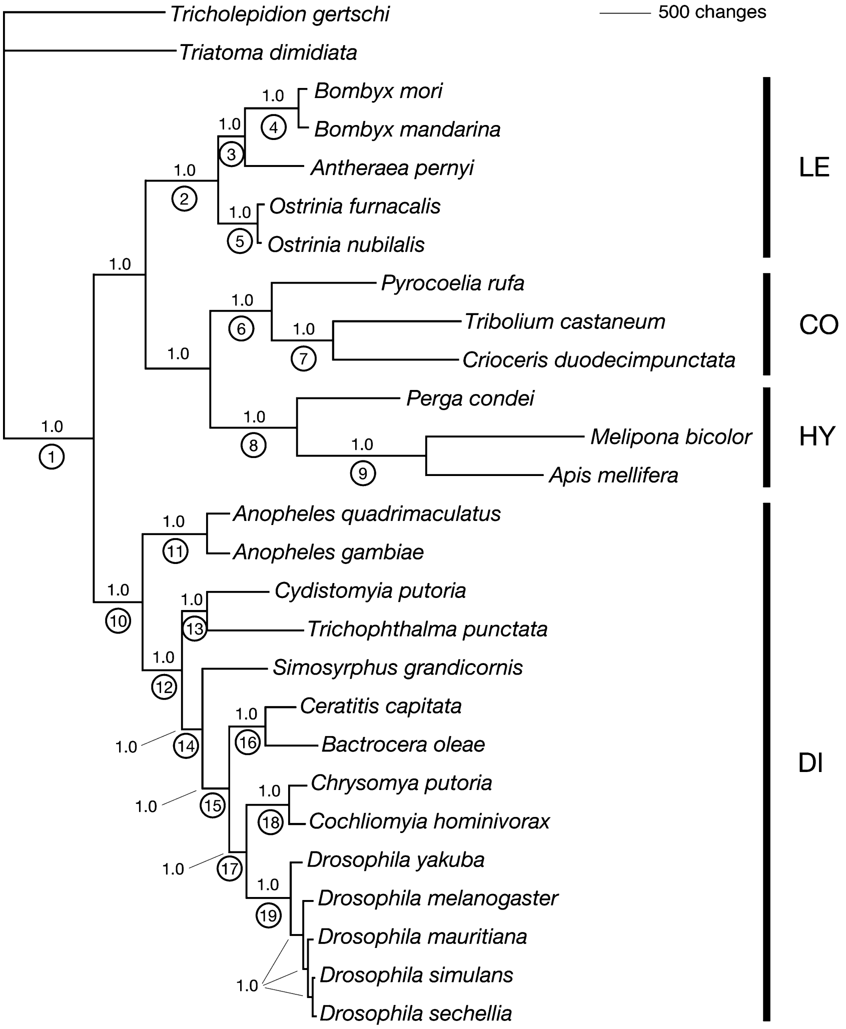

**Fig. 3.** Bayesian phenogram inferred from the ALL-DNA12 dataset. Posterior probabilities are indicated above each node. Numbers below each node (circled) are common across datasets; support levels across different datasets are given in Tables S5 and S6. CO, Coleoptera; DI, Diptera; HY, Hymenoptera; LE, Lepidoptera.

than Drosophilidae + Calliphoridae as in ALL and PCG analyses. It is possible that the tRNA genes are simply not variable enough to accurately capture mid-level dipteran relationships on their own, with stem sites supporting deep-level interordinal relationships and loop sites supporting shallow relationships. Within Drosophilidae, the TRAN datasets are in perfect concordance with accepted relationships within the drosophilids (Russo *et al.*, 1995; O'Grady & Kidwell, 2002; Lewis *et al.*, 2005; Symons & Wertheim, 2005) and with our larger PCG and ALL datasets.

Comparisons of how individual genes or sets of genes perform in isolated analyses, such as those described, fail to account for hidden support (*sensu* Gatesy *et al.*, 1999), whereby genes which, in isolation, support different topologies from those of combined analyses nonetheless actually support the combined analysis topology. The favoured method for assessing hidden support is comparison of partitioned Bremer support (= decay indices) across the combined topology. Partitioned Bremer values for each gene are given in Table 4. Across the genes, we see a typical mosaic of positive and negative values without a clear trend of particular genes driving the analysis or, conversely, consistently conflicting with the combined analysis. The one exception is the *l-rrna* gene, which is negative for all but the most derived relationships. This is not reflected in the *s-rrna* gene, which is positive for all but a single node (Tabanidae + Nemestrinidae), making it, together with *cox2*, the best predictor of combined topology in the mt genome. This suggests that the discrepancies between the RIBO and ALL datasets are driven either by the large ribosomal subunit, or support in the small subunit for the combined topology is very well hidden indeed. The tRNA genes are highly concordant with the combined topology, as is expected given the similarity in recovered topologies between the ALL and TRAN datasets. The only significantly negative value is

**Table 4.** Partitioned Bremer supports for the ALL-DNA12 dataset. Numbers and letters refer to the nodes labelled in Fig. 1.

| Node | *atp6* | *atp8* | *cox1* | *cox2* | *cox3* | *cytB* | *nadh1* | *nadh2* | *nadh3* | *nadh4* | *nadh4L* | *nadh5* | *nadh6* | *s-rrna* | *l-rrna* | tRNAs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Holometabola | 1 | 4 | 5.5 | 6 | 3.5 | −0.5 | 3 | −5 | −4 | 12.5 | 1.5 | 6 | 2.5 | 48.5 | 29 | 41.5 |
| 2. Coleoptera | 2 | 7 | −8 | −6 | −14 | −13 | 3 | 9 | 0 | 12 | −3 | 35 | −1 | 21 | 1 | 3 |
| 3. *Tribolium* + *Crioceris* | 1 | −1 | 5 | −2 | 2 | 3 | 6 | 1 | −1 | 6 | −1 | 1 | −6 | 12 | −15 | 5 |
| Diptera + (Lepidoptera + Hymenoptera) | −4 | 7 | 3 | 2 | −6 | −7 | −7 | −2 | 1 | 0 | −5 | 8 | −8 | 28 | 4 | −13 |
| Lepidoptera + Hymenoptera | −4 | 7 | 3 | 2 | −6 | −7 | −7 | −2 | 1 | 0 | −5 | 8 | −8 | 28 | 4 | −13 |
| 4. Lepidoptera | 8 | 0 | 19 | 5 | 12 | 10 | 15 | 7 | 4 | 15 | 7 | 42 | 19 | 27 | 50 | 54 |
| 5. *Antheraea* + *Ostrinia* | 1 | 3 | −3.5 | −3 | −2 | −7 | 3.5 | −9 | 1 | −3.5 | −0.5 | 10 | 1 | 21 | −5.5 | 0.5 |
| 6. *Bombyx* | 13 | 3 | 14 | 3 | 10 | 20 | 20 | 22 | 8 | 11 | 7 | 28 | 19 | 19 | 33 | 32 |
| 7. *Ostrinia* | 15 | 4 | 9 | 2 | 11 | 11 | 22 | 4 | 4 | 40 | 10 | 39 | 22 | 12 | 45 | 37 |
| 8. Hymenoptera | 15 | 9 | 8 | 5 | 1 | −15 | 14 | 9 | −5 | 18 | 9 | 19 | −5 | 27 | −26 | 14 |
| 9. Apidae | 40 | 0 | 63 | 23 | 51 | 42 | 70 | −8 | 26 | 79 | 14 | 121 | 20 | 37 | 8 | 52 |
| 10. Diptera | 0 | 7 | 6 | 4 | −2 | 1 | 0 | 1 | 1 | 8 | −4 | 23 | −6 | 30 | 19 | 4 |
| 11. ‘nematocera’ (= *Anopheles*) | 11 | 5 | 13 | 0 | 12 | 40 | 16 | 23 | 12 | 24 | 12 | 63 | 19 | 59 | 65 | 54 |
| 12. Brachycera | −8 | 4 | −8 | −2 | 6 | 2 | 8 | −2 | 0 | 1 | 10 | 24 | 2 | 17 | 4 | −2 |
| 13. Tabanidae + Nemestrinidae | −3 | 4 | 3 | 6 | −5 | −6 | −7 | 2 | 4 | −1 | −1 | 16 | −7 | 20 | 24 | −6 |
| 14. Cyclorrhapha | 1 | −2 | 1 | 0 | 2 | −2 | 4 | 6 | 4 | −4 | 1 | −3 | 4 | 9 | 11 | −11 |
| 15. Schizophora | −3 | 8 | 7 | 10 | −6 | −1 | −9 | −3 | 0 | 11 | −6 | 6 | −9 | 29 | 2 | −2 |
| 16. Tephritidae | 8.5 | 10 | 13 | 11 | 0.5 | 0 | 8 | 28.5 | 7.5 | 14 | 1.5 | 26 | −1 | 46.5 | 51.5 | 3.5 |
| 17. Drosphilidae + Calliphoridae | 1 | −5 | −1 | 1 | 4 | 1 | 0 | −1 | 1 | 5 | 0 | 6 | 6 | −3 | 4 | −1 |
| 18. Calliphoridae | 14 | 0 | 10 | 3 | 12 | 16 | 20 | 36 | 8 | 31 | 7 | 51 | 24 | 10 | 44 | 10 |
| 19. *Drosophila* | 5 | 1 | 3 | 10 | 13 | 14 | 14 | 33 | 2 | 39 | 8 | 44 | 21 | 15 | 26 | 27 |
| A. Node B + *D. melanogaster* | 0 | 1 | 2 | 0 | 2 | 6 | 1 | 2 | 2 | 3 | 0 | 8 | 3 | 7 | 10 | 7 |
| B. Node C + *D. mauritania* | 0 | 0 | 3 | 1 | 1 | 1 | 0 | 3 | 1 | 0 | 0 | 5 | 1 | 2 | 1 | 2 |
| C. *D. simulans* + *D. sechellia* | 4 | 0 | 2 | 2 | −1 | 0 | 1 | 5 | 0 | 5 | 1 | 4 | 4 | 3 | 6 | 1 |

for the monophyly of Cyclorrhapha as a result of the divergent tRNAs in the syrphid, which groups with strong support with *Anopheles*. It is also notable that the tRNA dataset frequently provides some of the strongest support for ordinal monophyly, for Holometabola and for other deep nodes, suggesting that it may be useful in resolving interordinal relationships within Insecta.

A second method of testing whether genes should be included in combined analyses is treewide assessment of homoplasy, the rationale being that highly homoplasious genes may decrease the resolution by more than is gained from the increased dataset size. A comparison of the relative contributions of each gene partition to the overall tree length is given in Table 5. Differences between protein coding genes are negligible. The ribosomal genes are less homoplasious than expected in the DNA123 and PROT analyses, but not in the DNA12 analyses, suggesting that third codon positions may be a more important source of homoplasy than the ribosomal genes. The tRNA genes are consistently the least homoplasious gene partition and the magnitude of their influence is greatly increased for the PROT analyses, suggesting that, with the loss of information due to recoding genes as amino acids, the tRNAs take on greater significance. There is no justification for the exclusion of any mt gene and there is a very strong argument for the inclusion of at least the tRNA genes.

In addition to partitioning by gene, we also examined the effects of partitioning within genes using the datasets DNA123, DNA12 and PROT, which should represent progressively more conservative analyses which sacrifice variability to reduce homoplasy. Comparison of the three approaches demonstrates the principal differences related to interordinal relationships (particularly for the MP analyses) and differing levels of nodal support within the orders. The DNA123 datasets tended to have higher nodal support than either the DNA12 or PROT datasets for common nodes, suggesting that the signal in the third codon position remains important despite high levels of homoplasy (cf. Table 3, the DNA123 trees are almost twice the length of the DNA12 trees). This trend extended to quite deep nodes, suggesting that the third codon positions do not become saturated quite as quickly as commonly believed (e.g. Lin & Danforth, 2004), although differences between interordinal relationships may be attributable to saturation. Further, the reduced nodal support and modest topological differences in some analyses between the DNA12 and PROT datasets conflicts with the widely held, but to the best of our knowledge never substantiated, notion that all differences between trees based on DNA vs. amino acids are due to third codon position variation. For ML and BA analyses, the outgroup hemipteran *Triatoma* tended to group within the ingroup close to Coleoptera in the PCG-DNA12 and PCG-PROT datasets. This did not occur in the ALL or PCG-DNA123 datasets,

**Table 5.** Relative contributions made by each gene to the whole phylogeny, ALL and PCG datasets. Differences in proportional length of each aligned gene vs. tree steps and informative sites only vs. tree steps. Negative values indicate that a gene's contribution to the overall tree length is less than expected based on its aligned length alone.

| | ALL-DNA123 | | ALL-DNA12 | | ALL-PROT | |
|---|---|---|---|---|---|---|
| Gene | Total | Informative | Total | Informative | Total | Informative |
| *atp6* | 0.20% | 0.06% | −0.30% | −0.21% | 0.07% | −0.11% |
| *atp8* | 0.36% | 0.16% | 0.47% | 0.20% | 0.75% | 0.54% |
| *cox1* | −1.48% | 0.44% | −3.97% | −0.26% | −2.83% | −0.32% |
| *cox2* | −0.07% | 0.03% | −0.66% | −0.04% | −0.10% | −0.12% |
| *cox3* | 0.07% | 0.07% | −0.81% | −0.31% | −0.06% | −0.05% |
| *cytB* | −0.01% | 0.18% | −1.29% | −0.22% | −0.31% | 0.02% |
| *nadh1* | −0.52% | −0.01% | −1.13% | −0.27% | 0.55% | 0.78% |
| *nadh2* | 1.89% | 0.24% | 2.22% | 0.22% | 3.64% | 1.95% |
| *nadh3* | 0.56% | 0.32% | 0.40% | 0.31% | 0.84% | 0.60% |
| *nadh4* | 1.10% | 0.39% | 0.62% | 0.16% | 2.51% | 1.36% |
| *nadh4L* | 0.13% | −0.06% | 0.25% | 0.00% | 0.73% | 0.43% |
| *nadh5* | 1.44% | 0.29% | 1.04% | 0.26% | 3.36% | 2.26% |
| *nadh6* | 1.38% | 0.50% | 1.61% | 0.53% | 2.40% | 1.49% |
| *l-rrna* | −0.91% | −0.44% | 2.29% | 0.88% | −2.46% | −2.36% |
| *s-rrna* | −0.47% | −0.54% | 1.40% | 0.06% | −1.24% | −1.90% |
| tRNAs | −3.68% | −1.63% | −2.15% | −1.31% | −7.86% | −4.57% |
| Gene | PCG-DNA123 | | PCG-DNA12 | | PCG-PROT | |
| *atp6* | −0.11% | −0.11% | −0.33% | −0.36% | −0.95% | −0.98% |
| *atp8* | 0.35% | 0.14% | 0.81% | 0.34% | 0.97% | 0.65% |
| *cox1* | −2.75% | 0.19% | −5.86% | −0.46% | −7.26% | −1.68% |
| *cox2* | −0.46% | −0.15% | −0.90% | −0.10% | −1.28% | −1.01% |
| *cox3* | −0.35% | −0.12% | −1.11% | −0.54% | −1.39% | −1.00% |
| *cytB* | −0.66% | −0.10% | −1.81% | −0.43% | −2.44% | −1.29% |
| *nadh1* | −1.21% | −0.27% | −1.54% | −0.46% | −0.84% | 0.07% |
| *nadh2* | 1.75% | −0.10% | 3.64% | 0.23% | 4.13% | 1.37% |
| *nadh3* | 0.51% | 0.29% | 0.71% | 0.47% | 0.75% | 0.44% |
| *nadh4* | 0.64% | 0.08% | 1.25% | 0.17% | 1.98% | 0.37% |
| *nadh4L* | 0.00% | −0.16% | 0.44% | −0.03% | 0.68% | 0.25% |
| *nadh5* | 0.90% | −0.13% | 2.06% | 0.40% | 2.65% | 1.31% |
| *nadh6* | 1.42% | 0.45% | 2.62% | 0.78% | 2.99% | 1.50% |

ALL, all gene partitions; DNA123, nucleotide sequences, all codon positions; DNA12, nucleotide sequences, first and second positions only; PCG, protein coding genes only; PROT, amino acid sequences.

which recovered holometabolan monophyly. Reducing the amount of data, through the elimination of third codon positions, recoding as amino acids or elimination of the rRNA and tRNA genes, apparently results in worse trees.

For the BA analyses, there is a related issue of how best to partition the data for the fastest and most accurate estimation of phylogeny. We performed duplicate analyses of the ALL-DNA123, ALL-DNA12, PCG-DNA123 and PCG-DNA12 datasets using three partitioning approaches – partitioning by gene (GP), by codon (CP) and by codons within each gene (CGP) – resulting in thirteen, three and thirty-nine partitions for the protein coding genes (for the ALL analyses, the rRNA and tRNA genes were included as an additional three partitions). Topologically, most differences were at the interordinal level. Partitioning had a greater influence on the DNA123 than the DNA12 datasets. In both ALL and PCG, GP favoured (Lepidoptera + Diptera) + (Coleoptera + Hymenoptera), whereas both CP and CGP favoured Diptera + (Lepidoptera + (Coleoptera + Hymenoptera)). The posterior probabilities for these alternative relationships were between 0.97 and 1.0 for ALL and 0.62 and 0.78 for PCG datasets, indicating that the partitioning strategy influences topology and that conflicting topologies can each enjoy significant support under different strategies. By contrast, the DNA12 topologies were identical for the three partitioning strategies in the ALL datasets and differed only in the location of *Triatoma* in the PCG datasets. This suggests that differences between the evolutionary dynamics of the first and second vs. third codon positions are more significant than differences between different mt genes. We calculated Bayes factors (Nylander *et al.*, 2004) comparing the three partition approaches (Table 6), showing that, for the DNA123 datasets, the difference between GP and CP was much greater than that between CP and CGP. For the ALL-DNA12 datasets, the difference between CP and CGP was much greater than that between GP and CP; however, the opposite was the case for PCG-DNA12 datasets. The magnitudes of each Bayes factor comparison were also extremely

**Table 6.** Bayes factors for comparison of partition strategy of protein coding genes.

| | ALL-DNA123 | ALL-DNA12 | PCG-DNA123 | PCG-DNA12 |
|---|---|---|---|---|
| Gene partitions (13) | −193964.88 | −120417.77 | −158292.09 | −82926.63 |
| Codon partitions (3) | −191348.76 | −120291.89 | −149817.65 | −78794.63 |
| Codon & gene partitions (39) | −189742.81 | −119137.02 | −147642.29 | −77579.68 |
| Bayes factor gene vs. codon | 2616.12 (very strong) | 125.88 (very strong) | 8474.44 (very strong) | 4132.00 (very strong) |
| Bayes factor codon vs. codon & gene | 1605.95 (very strong) | 1154.87 (very strong) | 2175.36 (very strong) | 1214.95 (very strong) |

ALL, all gene partitions; DNA123, nucleotide sequences, all codon positions; DNA12, nucleotide sequences, first and second positions only; PCG, protein coding genes only.

high, indicating highly significant differences. Collectively, the topological comparisons and the Bayes factors indicate that the correct partitioning strategy is very important to BA analyses of mt genome data, and that partitioning by codon (with or without additional gene-based partitions) is the most biologically relevant way to partition these data.

### *Reliability of mitochondrial genome data*

We can determine the relative timescales or ranges of molecular divergence over which mt genome data can be considered a reliable phylogenetic marker. For the ALL-DNA12 datasets, which had the best resolution, best nodal support and the least artefacts, relationships were consistent within each order. It is unclear whether these data are reliable at the interordinal level; however, as Diptera is the order for which taxonomic coverage is the most inclusive, we set the upper reliable bound at the divergence of Culicidae from the remaining dipterans. The lower bound is set by relationships within the genus *Drosophila*. All our analyses recovered the same branching pattern within *Drosophila*, a pattern which is consistent with the current phylogenetic understanding of this genus, and so we conclude that mt genomes are reliable down to at least the resolution of species within this genus. Current fossil and molecular clock estimates for these two divergences are as follows: Culicidae from the remaining Diptera in the early Triassic, 250–230 million years ago (Blagoderov *et al.*, 2002; Labandeira, 2005), and *D. simulans* + *D. sechellia* at less than one million years ago (Russo *et al.*, 1995; Tamura *et al.*, 2004). Alternatively, the calculation of divergences across these taxa (Table S7; ‘Supplementary material’) using corrected ML pairwise distances showed that these distances ranged from 0.01 (*D. simulans* vs. *D. sechellia*) to 0.47 (*A. quadrimaculatus* Culicidae vs. *T. punctata* Nemestrinidae). mt genome data are therefore a reliable phylogenetic marker over an extremely broad timescale (at least 1–200 million years ago) and across wide molecular divergences (0.1–0.47 ML distances).

## Discussion

### *Dipteran phylogeny and mitochondrial genomics*

Dipteran relationships are resolved robustly by mt genomes. Nodal support and topology are broadly consistent across analytical methods and are largely congruent with previous phylogenetic analyses of the order. Each of the four dipteran families represented by multiple exemplars (Culicidae, Tephritidae, Calliphoridae and Drosophilidae) are monophyletic. Species relationships within Drosophilidae are consistent with previous analyses based on multiple molecular markers (Russo *et al.*, 1995; O’Grady & Kidwell, 2002; Remsen & O’Grady, 2002; Lewis *et al.*, 2005; Symons & Wertheim, 2005) or morphology (Grimaldi, 1990; Schawaroch, 2002). Deeper relationships between dipteran families are also largely consistent with previous analyses (Yeates & Wiegmann, 2005) and recover a monophyletic Brachycera, Cyclorrhapha and Schizophora. The only inconsistency with previous work concerns the monophyly of the acalyptrates and placement of the Nemestrinidae (Fig. 4).

Our phylogeny strongly supports the sister group relationship, Nemestrinidae + Tabanidae. It is highly doubtful, however, that this is the actual sister group relationship, but rather the result of limited taxon sampling within the ‘basal brachycera’. Our result is more likely indicative of a close relationship between Nemestrinidae and one of the non-muscomorph brachyceran infraorders. Currently, Brachycera is divided into four infraorders, Xylophagomorpha, Stratiomyomorpha, Tabanomorpha and Muscomorpha (Yeates, 2002), with some evidence that the first three infraorders form a monophyletic group (Yeates, 2002; Wiegmann *et al.*, 2003; Yeates & Wiegmann, 2005). Nemestrinidae usually is considered part of Muscomorpha *sensu* Woodley (1989) (including Empidoidea, Asiloidea and Cyclorrhapha), although they have been included in Tabanoidea (Colless & McAlpine, 1991). Woodley (1989) proposed three Muscomorpha synapomorphies: reduction of antennal flagellomeres to four segments or less, loss of tibial spurs and female cerci composed of a single segment. Sinclair (1992), in a review of larval mandible characters, found that this character set neither added new support nor contradicted the monophyly of Muscomorpha. Two additional synapomorphies of Muscomorpha were proposed by Sinclair *et al.* (1994) from studies of male genitalia: base of the epandrium articulated on the gonocoxites (character 12) and gonostyli which move obliquely or in a dorsoventral direction (character 13). The first character state occurs also in Tabanidae and Xylomyidae, but its presence in these families is considered the result of convergent evolution (Sinclair *et al.*, 1994; Yeates, 2002). The second character, direction

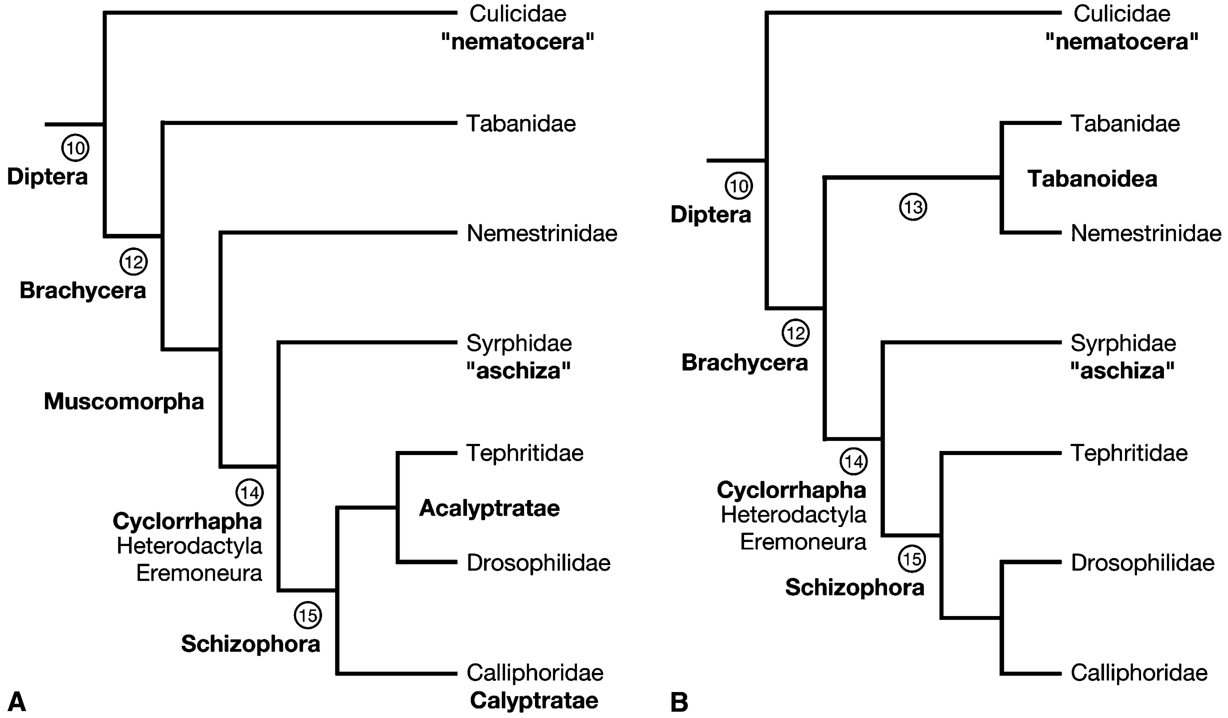

**Fig. 4.** Summarized relationships between dipteran families with major groups named. (A) Consensus of previous studies (redrawn after Yeates & Wiegmann, 2005). (B) This study. Note that Tabanoidea is *sensu* Colless & McAlpine (1989), Muscomorpha is *sensu* Woodley (1989) and Acalyptratae/Calyptratae are *sensu* McAlpine (1991). Numbers below each node (circled) are common across datasets (refer to other figures, Table 4 and supplementary material Tables S3–S6).

of gonostyli articulation, was scored incorrectly for Nemestrinidae, in which movement is horizontal as in 'nematocera', Xylophagomorpha, Stratiomyomorpha and Tabanomorpha (Yeates, 1994, 2002). Griffiths (1994) noted that each synapomorphy of Muscomorpha proposed by Woodley (1989) occurred also in Tabanomorpha. In Griffiths' (1994) opinion, evaluation of an alternative placement within Tabanomorpha, as advocated by Nagatomi (1992), required detailed morphological studies of the male genitalia in Nemestrinidae, as an existing study by Mackerras (1925) did not allow ready homologization of structures between nemestrinids and tabanomorphs. A contemporary detailed study of bombyliid phylogeny (Yeates, 1994), including a re-evaluation of nemestrinid morphology, does not comment on their possible homologies with tabanomorphs, but nemestrinids do not group with the included tabanomorph outgroup family Rhagionidae. Yeates (2002), in the first large-scale cladistic analysis of the group, found strong morphological support for the monophyly of Muscomorpha with Nemestrinidae, the sister group of Heterodactyla (the remaining muscomorphs). Six synapomorphies were found for muscomorphan monophyly, five of which are refinements of Woodley (1989) and Sinclair *et al.* (1994).

(1) Character 27: reduction of antennal flagellomeres to less than four.
(2) Character 53: foretibial spurs absent.
(3) Character 54: hindtibial spurs absent.
(4) Character 62: epandrium articulates on gonocoxites.
(5) Character 80: lateral aedeagal apodemes large and external.
(6) Character 98: female cerci one segmented.

However, echoing the concerns of Griffiths (1994), Yeates (2002) noted that most of these characters are at least partly homoplasious; antennal flagellomere reduction occurs in non-muscomorphan Brachycera, foretibial spurs are also absent in Pantophthalmidae, Xylomyidae and Stratiomyidae, hindtibial spurs are also absent in Stratiomyidae, articulated epandria are also found in Tabanidae and Xylomyidae, and females with one segmented cerci are also found in some Tabanidae and Athericidae. Finally, the novel muscomorphan synapomorphy, lateral aedeagal apodemes, has been interpreted in a wide variety of ways and is highly variable across Brachycera (Griffiths, 1994; Sinclair *et al.*, 1994). Most recently, Krzeminski & Krzeminski (2003) suggested that, on the basis of similarities of fossil specimens to xylophagomorphs and tabanomorphs, the traditional placement of Nemestrinidae within Muscomorpha should be reassessed.

Only one molecular study considering brachyceran relationships also includes a representative of the Nemestrinidae. Wiegmann *et al.* (2003), in a combined analysis of the large rRNA subunit (LSU-rRNA = 28S-rRNA) and morphology (the Yeates, 2002 matrix), found reasonable support (82% bootstrap) for the monophyly of Muscomorpha with Nemestrinidae as its earliest diverging branch. However, when morphology was excluded, the same topology was recovered but the nodal support became insignificant (less than 50% bootstrap support in MP and a posterior probability of 0.76 in BA analyses).

The question of the monophyly of Muscomorpha (*sensu* Woodley, 1989) is thus contentious. Morphological evidence favours the inclusion of nemestrinids in Muscomorpha, but most synapomorphies for this relationship require the postulation of convergent evolution of similar structures in other brachyceran infraorders. It is equally plausible that nemestrinids actually are related to one of these other groups, with the similarities of nemestrinids to heteroneurans being convergent. The only other molecular phylogenetic study on this question is clearly highly influenced by the morphological data partition, as LSU-rRNA phylogenies alone do not recover significant support for muscomorphan monophyly (Wiegmann *et al.*, 2003). By contrast, we found Tabanidae + Nemestrinidae in almost every analysis performed, with the relationship significantly supported in almost all cases. The only instance in which Nemestrinidae + Cyclorrhapha was found was the MP-TRAN dataset, without significant support. Additional research using mt genome phylogenies to test the monophyly and relationships of the brachyceran infraorders will probably be highly informative.

The second area in which our phylogeny differs from traditional ideas of dipteran relationships is in recovering a paraphyletic Acalyptratae. Acalyptrates are recognized as a monophyletic group in most major synthetic treatments (McAlpine, 1989; Yeates & Wiegmann, 1999, 2005), despite the acknowledgement that: 'The search for convincing synapomorphies uniting acalyptrate families has been difficult' (Yeates & Wiegmann, 1999: 416). The nominate character of the group, reduction of the calypter, is difficult to interpret across many schizophoran families as a result of its wide variation in both Acalyptratae and Calyptratae (= Muscoidea). Griffiths (1972) argued against Hennig's (1958) conception of acalyptrate monophyly on the basis of calypter variability, and stated that the other proposed character uniting acalyptrates, reduction of the pupal prothoracic spiracular horn, was both homoplastic and understudied in the group. Although arguing strongly against the monophyly of Acalyptratae, and also proposing calypter variability as useless for identification as a result of the inability of calypter morphology to even divide up schizophoran families, no proposal was made by Griffiths (1972) on how the five superfamilies within Schizophora were interrelated. By contrast, McAlpine (1989) strongly supported acalyptrate monophyly and proposed fourteen synapomorphies for the group. These proposals are yet to be tested extensively, and comprehensive phylogenetic treatment of the whole Schizophora to resolve the status of Acalyptratae has not been attempted (Yeates & Wiegmann, 1999; 2005).

In our opinion, the majority of recent molecular phylogenetic analyses of Diptera have not included sufficient samples for a confident resolution of the acalyptrate question. In a study of lower cyclorrhaphan relationships, principally the relationships of aschizan families to Schizophora, using LSU-rRNA, Collins & Wiegmann (2002) found a paraphyletic Acalyptratae, Otitidae + (Lauxaniidae + (Sepsidae + Muscidae)); sampling, however, was limited to a single species from each of just four families, a gross under-representation of schizophoran diversity. In Wiegmann *et al.* (2003), Acalyptratae and Calyptratae were represented by one species each, which were, unsurprisingly, sister groups. Studying Eremoneura, Moulton & Wiegmann (2004) included four schizophoran species representing four families, two each for calyptrates and acalyptrates, and recovered monophyly for both groups. The only study with substantial familial coverage within acalyptrates (Han & Ro, 2005) included fourteen of the sixty-two recognized families and seven of eleven superfamilies following the classification of Colless & McAlpine (1991). Their result was similar to that reported here, acalyptrate paraphyly with drosophilids grouping with calliphorids to the exclusion of tephritids, which is unsurprising given that this study exclusively used mt genes (*s-rrna*, *l-rrna* and *cox2*). Three previous mt genomic phylogenies have yielded a range of results on the question of schizophoran relationships. Nardi *et al.* (2001) found the same relationships as ours, using all the protein coding genes. Reanalysis (Nardi *et al.*, 2003a) favoured Calliphoridae + Tephritidae, but omitted nine of the thirteen protein coding genes and found non-significant support for any resolution within Schizophora. Support for acalyptrate monophyly also came from Junqueira *et al.* (2004), but five of the thirteen protein coding genes were omitted from this analysis. This diversity of results from similar primary data highlights the dangers of arbitrary data disposal (discussed further below and cf. Cameron *et al.*, 2004). Current mt genome sequences are inadequately diverse (two of sixty-two families) to be considered sufficient to address the question of acalyptrate monophyly, but the strength of nodal support for most relationships within Diptera, and the consistency of the recovered relationships with respect to analytical criteria, suggest that mt genome data may be an excellent avenue of investigation for addressing long-standing contentious issues in dipteran phylogeny, of which acalyptrate interrelationships are probably the most significant.

### *Methodological approaches to mitochondrial phylogenomics of insects*

Over a dozen studies have used the mt genome for insect phylogenetics on divergence scales as fine as between strains of *Drosophila simulans* (Ballard, 2000b) to as broad as deep-level hexapod phylogeny (Nardi *et al.*, 2003a). A wide variety of analytical approaches have been taken, but comparisons between approaches for common datasets have not been extensively performed. Exceptions include examinations of the effects of gene exclusion and optimality criteria (Stewart & Beckenbach, 2003; Kim *et al.*, 2005), of data coding strategies (Cameron *et al.*, 2004) and of taxon exclusion (Castro & Dowton, 2005). A key feature of each analysis is the focus on the effects of varying analytical regimes on insect interordinal relationships. As the rate at which mt genomes are being sequenced increases, it is worthwhile to continue and extend these efforts, particularly concerning how methodological approaches may affect the accuracy of inferring intraordinal relationships. For that reason, now that sufficient representatives are available to examine the relationships of the major dipteran clades, we chose to address some of the factors that may affect mt genome phylogenies significantly – optimality criteria, data inclusion/exclusion, data transformations and partitioning strategies.

Optimality criteria are a perennial source of argument about the accuracy of phylogenetic studies. Fortunately, optimality criteria do not seem to have a very significant effect on intraordinal relationships. For common datasets, the topology within Diptera recovered by different criteria is largely the same. The major differences lie in the strength of nodal support, with MP bootstrapping consistently giving the weakest and BA posterior probabilities the strongest support for nodes which are common across optimality criteria. This effect has repeatedly been noted (Leaché & Reeder, 2002; Whittingham *et al.*, 2002; Cameron *et al.*, 2004) in comparisons of bootstrapping with BA posterior probabilities. Bootstrapping is known to be sensitive to the level of homoplasy in a dataset (Hillis & Bull, 1993). Given that the consistency index of the datasets analysed here never increased above 0.6 (Table 5), it is likely that bootstrapping underestimates nodal support for mt genome data. We have used a cut-off of 70% bootstrap support for nodal significance in line with previous estimations based on smaller, less noisy, simulated datasets (Hillis & Bull, 1993); however, it is possible that significance levels may need to be re-examined for the much larger and noisier, phylogenomic datasets currently available. Conversely, we are suspicious that BA posterior probabilities overstate nodal significance, particularly in the light of our finding that the same dataset, when partitioned in different ways, gave alternative topologies for which every node received posterior probabilities of 1.0 (see the fuller discussion below). Despite the similar topological results from different optimality criteria, we still believe that it is valuable to perform replicate analyses under different schemes as a way of evaluating upper (BA) and lower (MP) confidence bounds to recovered topologies.

The second major issue in phylogenetic analyses of mt genome sequences is data exclusion. Many studies have excluded one (e.g. Friedrich & Muqim, 2003), two (Lessinger *et al.*, 2000) or as many as nine (Nardi *et al.*, 2003a) of the thirteen protein coding genes from phylogenetic analysis, often citing difficulties of alignment as justification. In our opinion, these concerns are overstated. Reducing the amount of available data resulted in a loss of resolution and weakening of nodal support in the present study. The ALL datasets were usually better resolved, better supported and gave the most reasonable results in the light of previous phylogenetic work within Diptera. Levels of homoplasy did not vary greatly between the protein coding genes, which mirrors our previous findings for a much wider phylogenetic study of mt genomes across all arthropods (Cameron *et al.*, 2004). The pattern of Bremer support across the gene partitions did not suggest that any particular gene had a radically different history, and partitions which, in isolation (RIBO, TRAN), gave radically different topologies showed considerable levels of hidden support for the ALL topology. This is to be expected, as recombination is supposedly rare within mt genomes, the mt genes are evolving as a linked unit and any particular gene is unlikely to have an evolutionary history entirely at odds from the rest of the genome (Ballard & Rand, 2005). The only major reason for phylogenetic 'untrustworthiness' is therefore differing levels of noise between genes and, indeed, Saccone *et al.* (1999) have demonstrated that substitution rates may be related to the physical position of individual genes within the mt genome. Our results, however, do not suggest that there is any correlation between gene location and phylogenetic behaviour, as no gene is consistently for or against the topology (Fig. 1; Table 6). Further, the difficulty of aligning protein coding genes can be alleviated in several ways. The omission of highly divergent genomes that are not critical to the phylogenetic questions under consideration, such as, in this instance, the louse and sternorrhynch hemipterans, and better balanced taxon selections which include more than just a few representatives of included clades, make automated alignment more consistent. Secondly, the use of amino acid translations for alignments before back-translating for analysis results in more reasonable alignments than those generated by nucleotides alone.

Related to the question of eliminating protein coding genes is the lack of attention which has been devoted to rRNA and tRNA genes. Prior to this study, rRNA genes had only been included three times (Stewart & Beckenbach, 2003; Bae *et al.*, 2004; Kim *et al.*, 2005) and tRNAs once (Kim *et al.*, 2005) in phylogenetic analyses of insects. Kim *et al.* (2005) noted that, in comparisons of pairwise differences between taxa, the tRNAs had consistently lower distances than either the protein coding or rRNA genes. Similarly, we found that the tRNAs were much less homoplasious than expected from their combined length, and that the TRAN dataset gave comparable results to the ALL or PCG datasets for deep and shallow nodes. It should be noted, however, that we did not incorporate either paired site coding (Gillespie, 2004) within the MP analyses or paired site models, such as the doublet model, in MrBayes (Ronquist & Huelsenbeck, 2003); therefore, potentially, support from this partition may be inflated as informative stem sites are counted twice rather than once (Wheeler & Honeycutt, 1988). Despite this, tRNA genes have been needlessly overlooked as a result of untested assumptions regarding the rates of nucleotide evolution in mitochondria, i.e. that tRNA genes are generally too fast. By contrast, the rRNA genes probably warrant more caution in their use. The RIBO datasets gave tree topologies that were wildly incongruent with the other datasets and with previously accepted dipteran phylogenies, particularly for the ML and BA analyses, and the large subunit RNA was highly homoplasious. Kim *et al.* (2005) found that the rRNA genes had consistently much higher pairwise differences than the protein coding genes, suggesting that the rRNA genes may be very noisy. An additional difficulty with the use of the ribosomal genes is that they have not been consistently annotated across sequenced insect mt genomes. The large subunit is defined as all nucleotides between adjacent tRNAs, whereas the small subunit is similarly defined as all nucleotides between an adjacent tRNA and the origin of replication, which itself has not been experimentally defined for any insect other than *Drosophila* (Clayton, 1992). The development of functional methods of annotating mt rRNA genes based on secondary structure analyses, similar to those used to define tRNAs, is necessary before the rRNA genes can be confidently compared in phylogenetic analyses.‡

The third major area of methodological variation between insect phylogenetic studies of the mt genome is how sequence data are treated in the analysis. Three approaches have been investigated, two of which are examined in this study. The first is the translation of nucleotides to amino acids, which has the advantage of reducing the noise associated with the saturated nucleotide sequence, but introduces other potential sources of noise due to redundancy in the genetic code (cf. Inagaki *et al.*, 2004). This approach was one of the first used and has been widely adopted (e.g. Nardi *et al.*, 2001, 2003a; Friedrich & Muqim, 2003; Cameron *et al.*, 2004). The second approach has been to use nucleotides directly without translation, which has the advantage of using the maximal amount of data; however, justifiable concerns about the capacity of noise in saturated nucleotide datasets to overwhelm the signal remains. Nucleotide analyses have also been widely used, both in comparison with amino acid analyses (e.g. Cameron *et al.*, 2004; present study) and as stand-alone analyses (e.g. Bae *et al.*, 2004; Castro & Dowton, 2005; Kim *et al.*, 2005). The third and most recently proposed approach is the translation of portions of the nucleotide data into purines (A or G = R) or pyrimidines (C or T = Y) – R/Y coding – to reduce the effects of nucleotide compositional bias (Phillips & Penny, 2003). Delsuc *et al.* (2003) took this approach to reanalyse the data of Nardi *et al.* (2003a) supporting hexapod polyphyly, and claimed that R/Y coding supported a monophyletic Hexapoda. This result, however, was not even remotely significant ($P = 0.57$), and subsequent attempts to apply R/Y coding to questions of polyneopteran relationships resulted in an almost complete collapse of signal on recoding of nucleotides (Cameron *et al.*, 2006). In the light of the limited usefulness of R/Y coding, we have not tested again in this study. A fourth approach of combining data translated at multiple levels, such as nucleotides with amino acids or nucleotides with RY and amino acids, has yet to be applied to mt genome data (e.g. Agosti *et al.*, 1995).

The present study supported the use of nucleotides over amino acids for intraordinal phylogenies because, even though the topology within Diptera was largely the same for both data treatments, nodal support was higher for nucleotide datasets. This suggests that recoding nucleotide sequences as amino acids results in the loss of informative nucleotide variation, such as that due to silent substitutions at the third codon position. Concerns about third position saturation are probably overstated, as we saw little variation between the DNA12 and DNA123 datasets. Only once did inclusion of the third codon position result in an artefactual grouping, namely syrphids grouping within Schizophora in the ALL-DNA123 dataset in the ML analysis, and even then this relationship received limited nodal support. By contrast, for interordinal relationships, data treatment appears to be very significant, resulting in differing relationships between the holometabolan orders for the different data treatments and, particularly, for the outgroup hemipteran *Triatoma*, which tended to nest deep inside Holometabola in some treatments. In general, however, the use of amino acid sequence for alignment balances concerns about nucleotide saturation against loss of signal. We have found that aligning protein coding genes directly in programs such as Clustal results in very 'gappy' alignments and very poor trees (data not shown). The direct alignment of nucleotides using ClustalX (Thompson *et al.*, 1997) has also been tried by Bae *et al.* (2004) and Kim *et al.* (2005). These papers reported unusual results relating to the early branching patterns [e.g. outgroup chelicerates and crustaceans within Insecta and Orthoptera + Holometabola to the exclusion of Paraneoptera (Bae *et al.*, 2004); and a polyphyletic Orthoptera with ensiferans forming the earliest branch within insects (Kim *et al.*, 2005)]. This suggests that the approach of direct nucleotide alignments may be very sensitive to outgroup choice, particularly over long internodes. Interestingly, this mirrors the results of Cameron *et al.* (2004) relating to the influence of outgroup on deep-level insect mitogenomics. Amino acid alignment with back-translation to nucleotides produces the most reasonable results, but this procedure forces insertion–deletion events (indels) into neat three nucleotide sets, which is unrealistic biologically as it ignores the possibility of single or double nucleotide indels, both of which have been recorded (Beckenbach *et al.*, 2005) and inferred (Grant & D'Haese, 2004) for mt genes.

Although BA analysis is now a standard feature of mt phylogenomic studies of the insects (used in Nardi *et al.*, 2003a; Cameron *et al.*, 2004; Castro & Dowton, 2005), this is the first examination of the effect of data partitioning on BA analysis. Partitioning improves both the accuracy and speed of BA analyses, but over-partitioning results in inaccuracies similar to those encountered in ML analysis in which models are over-parameterized (e.g. Nylander *et al.*, 2004). It is also a potential area of concern because, as found here, the same dataset partitioned in different ways can support different topologies, with the differing nodes of each alternative topology supported by significant (0.9) to

‡A secondary structure model for the mt rRNA genes of *Apis* will be published shortly (Gillespie *et al.*, 2006), and its application to other insect groups will no doubt greatly improve the alignments of insect mt rRNA genes for phylogenetic studies.

absolute (1.0) posterior probabilities. We examined three possible ways of partitioning the protein coding genes – CP, GP and CGP – resulting in three, thirteen and thirty-nine partitions, respectively. Interestingly, differing partitioning strategies resulted in different topologies for the DNA123 datasets; however, these were only at the interordinal level and only significant for the ALL-DNA123 dataset. In addition, the trees from CP and CGP analyses were the same, whereas the GP trees differed for the DNA123 datasets, but all three analyses were basically the same for the DNA12 datasets. Allowing different parameter optimizations between genes was far less important than allowing them between codons, which is why analyses that included some form of codon partition were the same. The absence of this 'codon effect' for the DNA12 datasets shows that significant rate heterogeneity exists between first + second and third codon positions, which is to be expected given that the third codon position is typically mutationally fast as substitutions are usually silent. Although the partitioning strategy does not appear to have a significant effect on intraordinal topology, the highly significant differences in Bayes factors between the three approaches suggest that it is still important, especially if accurate calculation of branch lengths is important, such as for molecular clock calculations. From this perspective, partitioning by codon alone may not be enough. Partitioning by gene, however, has the potential drawback that some of the mt genes may be too small for accurate a priori determination of parameter values in programs such as ModelTest (it was for this reason that the tRNA genes were treated as a single partition in all analyses). For example, for *atp8*, ModelTest favoured the Jukes–Cantor model, an evolutionary model so simplistic as to be almost completely unrealistic, which may be the result of the small size of this gene (183 bp aligned, 61 bp for each codon partition). Therefore, it may be desirable to treat blocks of genes as partitions, particularly those that, in the vast majority of insect mt genomes, have partially overlapping coding regions, such as *atp8–atp6–cox3*, *nadh4L–nadh4* and *nadh6–cytB*.

## Conclusion

We have demonstrated the broad usefulness of mt genome data for resolving intraordinal relationships within insects by a pilot study of the best represented insect order, Diptera. We have also evaluated many of the analytical approaches that have been investigated for insect phylogenetic reconstruction, and can conclude that the following elements are important considerations in experimental/analytical design.

(1) The effect of optimality criteria on topology and tree support should be assessed by performance of replicate analyses because, even for regions of topological congruence between methods, nodal support can vary widely.

(2) Analyses should use as much data as possible. There is no evidence of significant incongruence between mt genes and, if alignment of divergence sequences remains a concern, taxon exclusion is a better approach (e.g. see Castro & Dowton, 2005 for the effects of excluding the highly divergent *Apis* and *Melipona* sequences).

(3) The rRNA and tRNA genes should be included in future analyses of insect relationships. There is considerable phylogenetic signal in these genes, adding almost 25% extra data sequenced but discarded from most genomic studies. Difficulties with the alignment of the rRNA genes, particularly *l-rrna*, can be improved by the adoption of more consistent functional genome annotation approaches.

(4) Sequence recoding as amino acids is unnecessary, reduces the phylogenetic signal and results in artefactual relationships, placing outgroup sequences in the ingroup. Concerns about the saturation of third codon positions can be more readily addressed by the inclusion/exclusion of this data partition. Recoding sequences as amino acids is, however, very useful in the alignment phase of an analysis.

(5) BA analyses are extremely sensitive to partition strategy. Partitions based on codons give different results to those based on genes, but both are very well supported. Bayes factors suggest that codon-based partitions are better than gene-based partitions; however, over-partitioning will probably result in poor parameter estimation by both ModelTest and MrBayes.

## Supplementary material

Tables S1–S7 listed in the text, which include primer sequences, nodal supports and calculation of pairwise divergences, are available online at www.blackwell-synergy.com under DOI reference doi: 10.1111/j.1365-3113.2006.00355.x

## Acknowledgements

We thank Brian Wiegmann (North Carolina State University) and David Yeates (Australian National Insect Collection) for their advice and support of this research. Mark Dowton (University of Wollongong) and Joseph Gillespie (University of Maryland) provided valuable comments on the manuscript. This study was supported by NSF grants DEB0120718 (MFW) and DEB0444972 (SLC, MFW).

## References

Agosti, D., Jacobs, D. & DeSalle, R. (1995) On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics*, **12**, 65–82.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bae, J.S., Kim, I., Sohn, H.D. & Jin, B.R. (2004) The mitochondrial genome of the firefly, *Pyrocoelia rufa*: complete DNA sequence, genome organization, and phylogenetic analysis with other insects. *Molecular Phylogenetics and Evolution*, **32**, 978–985.

Baker, R.H. & DeSalle, R. (1997) Multiple sources of character information and the phylogeny of the Hawaiian drosophilids. *Systematic Biology*, **46**, 654–673.

Ballard, J.W.O. (2000a) Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *Journal of Molecular Evolution*, **51**, 48–63.

Ballard, J.W.O. (2000b) Comparative genomics of mitochondrial DNA in *Drosophila simulans*. *Journal of Molecular Evolution*, **51**, 64–75.

Ballard, J.W.O. & Rand, D.M. (2005) The population biology of mitochondrial DNA and its phylogenetic implications. *Annual Review of Ecology, Evolution and Systematics*, **36**, 621–642.

Beard, C.B., Hamm, D.M. & Collins, F.H. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organisation and comparisons with mitochondrial sequences of other insects. *Insect Molecular Biology*, **2**, 103–124.

Beckenbach, A.T., Robson, S.K.A. & Crozier, R.H. (2005) Single nucleotide + 1 frameshifts in an apparently functional mitochondrial cytochrome b gene in ants of the genus *Polyrhachis*. *Journal of Molecular Evolution*, **60**, 141–152.

Blagoderov, V.A., Lukashevich, E.D. & Mostovski, M.D. (2002) Order Diptera Linné, 1758. *History of Insects* (ed. by A. P. Rasnitsyn & D. L. J. Quicke), pp. 227–240. Kluwer Academic Publishers, Dordrecht.

Boore, J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Research*, **27**, 1767–1780.

Boore, J.L. & Brown, W.M. (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics and Development*, **8**, 668–674.

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

Cameron, S.L., Barker, S.C. & Whiting, M.F. (2006) Mitochondrial genomics and the new insect order Mantophasmatodea. *Molecular Phylogenetics and Evolution*, **38**, 274–279.

Castro, L.R. & Dowton, M. (2005) The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae). *Molecular Phylogenetics and Evolution*, **34**, 469–479.

Caterino, M.S., Cho, S. & Sperling, F.A.H. (2000) The current state of insect molecular systematics: a thriving Tower of Babel. *Annual Review of Entomology*, **45**, 1–54.

Clary, D.O. & Wolstenholme, D.R. (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, **22**, 252–271.

Clayton, D.A. (1992) Transcription and replication of animal mitochondrial DNAs. *International Review of Cytology*, **141**, 217–232.

Coates, B.S., Sumerford, D.V., Hellmich, R.L. & Lewis, L.C. (2005) Partial mitochondrial genome sequences of *Ostrinia nubialis* and *Ostrinia furnicalis*. *International Journal of Biology Sciences*, **1**, 13–18.

Colless, D.H. & McAlpine, D.K. (1991) Diptera. The Insects of Australia (ed. by CSIRO), pp. 717–786. Melbourne University Press, Melbourne.

Collins, K.P. & Wiegmann, B.M. (2002) Phylogenetic relationships of the lower Cyclorrhapha (Diptera: Brachycera) based on 28S rDNA sequences. *Insect Systematics and Evolution*, **33**, 445–456.

Covacin, C., Shao, R., Cameron, S.L. & Barker, S.C. (2006) Extraordinary amounts of gene rearrangement in the mitochondrial genomes of lice (Insecta: Phthiraptera). *Insect Molecular Biology*, **15**, 63–68.

Crozier, R.H. & Crozier, Y.C. (1993) The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organisation. *Genetics*, **133**, 97–117.

Delsuc, F., Phillips, M.J. & Penny, D. (2003) Comment on 'Hexapod origins: Monophyletic or Paraphyletic'. *Science*, **301**, 1482d.

Dotson, E.M. & Beard, C.B. (2001) Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*. *Insect Molecular Biology*, **10**, 205–215.

Dowton, M. & Austin, A.D. (1999) Evolutionary dynamics of a mitochondrial rearrangement 'hotspot' in the Hymenoptera. *Molecular Biology and Evolution*, **16**, 298–309.

Dowton, M., Castro, L.R. & Austin, A.D. (2002) Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. *Invertebrate Systematics*, **16**, 345–356.

Dowton, M., Castro, L.R., Campbell, S.L., Bargon, S.D. & Austin, A.D. (2003) Frequent mitochondrial gene rearrangements at the hymenopteran nad3nad5 junction. *Journal of Molecular Evolution*, **56**, 517–526.

Flook, P.K., Rowell, C.H.F. & Gellissen, G. (1995a) The sequence, organisation and evolution of the *Locusta migratoria* mitochondrial genome. *Journal of Molecular Evolution*, **41**, 928–941.

Flook, P.K., Rowell, C.H.F. & Gellissen, G. (1995b) Homoplastic rearrangements of insect mitochondrial tRNA genes. *Naturwissenschaften*, **82**, 336–337.

Friedrich, M. & Muqim, N. (2003) Sequence and phylogenetic analysis of the complete mitochondrial genome of the flour beetle *Tribolium castanaeum*. *Molecular Phylogenetics and Evolution*, **26**, 502–512.

Gatesy, J., O'Grady, P. & Baker, R. (1999) Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, **15**, 271–313.

Gillespie, J.J. (2004) Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. *Molecular Phylogenetics and Evolution*, **33**, 936–943.

Gillespie, J.J., Johnston, J.S., Cannone, J.J. & Gutell, R.R. (2006) Characteristics of the nuclear (18S, 5.8S, 28S, and 5S) and mitochondrial (12S and 16S) rDNA genes of *Apis mellifera* (Insecta: Hymenoptera): Structure, organization, and retrotransposition. *Insect Molecular Biology* (in press).

Grant, T. & D'Haese, C.A. (2004) Insertions and deletions in the evolution of equal-length DNA fragments. *Cladistics*, **20**, 84.

Griffiths, G.C.D. (1972) *The Phylogenetic Classification of Diptera Cyclorrhapha with Special Reference to the Structure of the Male Postabdomen*. W. Junk N.V. Publishers, The Hague.

Griffiths, G.C.D. (1994) Relationships among the major subgroups of Brachycera (Diptera): a critical review. *Canadian Entomologist*, **126**, 861–880.

Grimaldi, D.A. (1990) A phylogenetic, revised classification of genera in the Drosophilidae (Diptera). *Bulletin of the American Museum of Natural History*, **197**, 1–139.

Han, H.-Y. & Ro, K.-E. (2005) Molecular phylogeny of the superfamily Tephritoidea (Insecta: Diptera): new evidence from the mitochondrial 12S, 16S and COII genes. *Molecular Phylogenetics and Evolution*, **34**, 416–430.

Hennig, W. (1958) Die Familien der Diptera Schizophora und ihre phylogenetischen Verwandtschaftsbeziehungen. *Beiträge zur Entomologie*, **8**, 505–688.

Hillis, D.M. & Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, **42**, 182–192.

Huelsenbeck, J.P. & Ronquist, F.R. (2001) MrBayes: Bayesian inference of phylogeny. *Biometrics*, **17**, 754–755.

Inagaki, Y., Simpson, A.G.B., Dacks, J.B. & Rojer, A.J. (2004) Phylogenetic artifacts can be caused by leucine, serine and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. *Systematic Biology*, **53**, 582–593.

Junqueira, A.C.M., Lessinger, A.C., Torres, T.T., Rodrigues da Silva, F., Vettore, A.L., Arruda, P. & Azeredo Espin, A.M.L. (2004) The mitochondrial genome of the blowfly *Chrysomya chloropyga* (Diptera: Calliphoridae). *Gene*, **339**, 7–15.

Kim, I., Cha, S.Y., Yoon, M.H., Hwang, J.S., Lee, S.M., Sohn, H.D. & Jin, B.R. (2005) The complete nucleotide sequence and gene organisation of the mitochondrial genome of the oriental mole cricket. *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae). *Gene*, **353**, 155–168.

Kluge, A.G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae: Serpentes). *Systematic Zoology*, **38**, 7–25.

Kristensen, N.P. & Skalski, A.W. (1999) Phylogeny and palaeontology. Handbuch der Zoologie Band IV Arthropoda: Insecta Part 35 Lepidoptera, Moths and Butterflies (ed. by N. P. Kristensen), pp. 7–25. Walter de Gruyter Publications, Berlin.

Krzeminski, W. & Krzeminski, E. (2003) Triassic Diptera: descriptions, revisions and phylogenetic relations. *Acta Zoologica Cracoviensia*, **46S**, 153–184.

Kumar, S., Tamura, K. & Nei, M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Labandeira, C.C. (2005) Fossil history and evolutionary ecology of Diptera and their associations with plants. *The Evolutionary Biology of Flies* (ed. by D. K. Yeates & B. M. Wiegmann), pp. 217–273. Columbia University Press, New York.

Leaché, A.D. & Reeder, T.W. (2002) Molecular systematics of the eastern fence lizard *Sceloporus undulates*: a comparison of parsimony, likelihood and Bayesian approaches. *Systematic Biology*, **51**, 44–68.

Lessinger, A.C., Junqueira, A.C.M., Lemos, T.A., Kemper, E.L., Rodrigues da Silva, F., Vettore, A.L., Arruda, P. & Azeredo Espin, A.M.L. (2000) The mitochondrial genome of the primary screwworm fly *Cochliomyia hominovorax* (Diptera: Calliphoridae). *Insect Molecular Biology*, **9**, 521–529.

Lessinger, A.C., Junqueira, A.C.M., Conte, F.F. & Azeredo Espin, A.M.L. (2004) Analysis of a conserved duplicated tRNA gene in the mitochondrial genome of blowflies. *Gene*, **339**, 1–6.

Lewis, D.L., Farr, C.L. & Kaguni, L.S. (1995) *Drosophila melanogaster* mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. *Insect Molecular Biology*, **4**, 263–278.

Lewis, R.L., Beckenbach, A.T. & Mooers, A.Ø. (2005) The phylogeny of the subgroups within the *melanogaster* species group: Likelihood tests on *CO1* and *COII* sequences and a Bayesian estimate of phylogeny. *Molecular Phylogenetics and Evolution*, **37**, 15–24.

Lin, C.P. & Danforth, B.N. (2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analysis of combined datasets. *Molecular Phylogenetics and Evolution*, **30**, 686–702.

Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964.

Mackerras, I.M. (1925) The Nemestrinidae (Diptera) of the Australasian Region. *Proceedings of the Linnean Society of New South Wales*, **50**, 489–561.

Maddison, W. & Maddison, D. (2003) *Macclade Version 4.06*. Sinauer Associates, Sunderland, MA.

McAlpine, J.F. (1989) Phylogeny and classification of the Muscomorpha. *Manual of Nearctic Diptera*, Vol. 3 (ed. by J. F. McAlpine & D. V. Wood), pp. 1397–1518. Research Branch, Agriculture Canada, Ottawa.

Mitchell, S.E., Cockburn, A.F. & Seawright, J.A. (1993) The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and genome organisarion. *Genome*, **36**, 1058–1073.

Moulton, J.K. & Wiegmann, B.M. (2004) Evolution and phylogenetic utility of CAD (rudimentary) among Mesozoic-aged Eremoneuran Diptera (Insecta). *Molecular Phylogenetics and Evolution*, **31**, 363–378.

Nagatomi, A. (1992) Notes on the phylogeny of various taxa of the orthorraphous Brachycera (Insecta: Diptera). *Zoological Science*, **9**, 843–857.

Nardi, F., Carapelli, A., Fanciulli, P.P., Dallai, R. & Frati, F. (2001) The complete mitochondrial DNA sequence of the basal hexapod *Tetrodontophora bielanensis*: evidence for heteroplasmy and tRNA translocations. *Molecular Biology and Evolution*, **18**, 1293–1304.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R. & Frati, F. (2003a) Hexapod origins: monophyletic or paraphyletic? *Science*, **299**, 1887–1889.

Nardi, F., Carapelli, A., Dallai, R. & Frati, F. (2003b) The mitochondrial genome of the olive fly *Bactrocera oleae*: two haplotypes from distant geographical locations. *Insect Molecular Biology*, **12**, 605–611.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P. & Nieves-Aldrey, J.L. (2004) Bayesian phylogenetic analysis of combined data. *Systematic Biology*, **53**, 47–67.

O'Grady, P.M. & Kidwell, M.G. (2002) Phylogeny of the subgenus *Sophophora* (Diptera: Drosphilidae) based on combined analysis of nuclear and mitochondrial sequences. *Molecular Phylogenetics and Evolution*, **22**, 442–453.

Phillips, M.J. & Penny, D. (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, **28**, 171–185.

Posada, D. & Crandall, K.A. (1998) ModelTest: Testing the best-fit model of nucleotide substitution. *Bioinformatics*, **14**, 817–818.

Remsen, J. & O'Grady, P.M. (2002) Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Molecular Phylogenetics and Evolution*, **24**, 249–264.

Rokas, A. & Holland, P.W.H. (2000) Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, **15**, 454–459.

Ronquist, F. & Huelsenbeck, J.P. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Russo, C.A.M., Takezaki, N. & Nei, M. (1995) Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution*, **12**, 391–404.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. & Reyes, A. (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, **238**, 195–209.

Schawaroch, V. (2002) Phylogeny of a paradigm lineage: the *Drosophila melanogaster* species group (Diptera: Drosophilidae). *Biological Journal of the Linnean Society*, **76**, 21–37.

Shao, R., Campbell, N.J.H. & Barker, S.C. (2001) Numerous gene rearrangements in the mitochondrial genome of the wallaby louse, *Heterodoxus macropus* (Phthiraptera). *Molecular Biology and Evolution*, **18**, 858–865.

Sinclair, B.J. (1992) A phylogenetic interpretation of the Brachycera (Diptera) based on the larval mandible and associated mouthpart structures. *Systematic Entomology*, **17**, 233–252.

Sinclair, B.J., Cumming, J.M. & Wood, D.M. (1994) Homology and phylogenetic implications of male genitalia in Diptera – Lower Brachycera. *Entomologica Scandinavica*, **24**, 407–432.

Sorenson, M.D. (1999) TreeRot, Version 2. Boston University, Boston, MA.

Spanos, L., Koutroumbas, G., Kotsyfakis, M. & Louis, C. (2000) The mitochondrial genome of the Mediterranean fruit fly, *Ceratitis capitata. Insect Molecular Biology*, **9**, 139–144.

Stewart, J.B. & Beckenbach, A.T. (2003) Phylogenetic and genomic analysis of the complete mitochondrial DNA sequence of the spotted asparagus beetle *Crioceris duodecimpunctata. Molecular Phylogenetics and Evolution*, **26**, 513–526.

Swofford, D.L. (2002) *Paup* Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, MA.

Symons, M.R.E. & Wertheim, B. (2005) The mode of evolution of aggregation pheromones in *Drosophila* species. *Journal of Evolutionary Biology*, **18**, 1253–1263.

Tamura, K., Subramanian, S. & Kumar, S. (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, **21**, 36–44.

Terry, M.D. & Whiting, M.F. (2005) Comparison of two alignment techniques within a single complex data set: POY versus Clustal. *Cladistics*, **21**, 272–281.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. (1997) The CLUSTAL–windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **24**, 173–216.

Vogler, A.P. (2005) Molecular systematics of Coleoptera: what has been achieved so far? *Handbuch der Zoologie, Band IV Arthropoda: Insecta Part 38 Coleoptera. Beetles* (ed. by N. P. Kristensen & R. G. Beutel), pp. 17–22. Walter der Gruyter Press, Berlin.

Wheeler, W.C. & Honeycutt, R.L. (1988) Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Molecular Biology and Evolution*, **5**, 90–96.

Whittingham, L.A., Slikas, B., Winker, D.W. & Sheldon, F.H. (2002) Phylogeny of the tree swallow genus *Tachycineta* (Aves: Hirundinidae) by Bayesian analysis of mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **22**, 430–441.

Wiegmann, B.M., Tsaur, S.-C., Webb, D.W., Yeates, D.K. & Cassel, B.K. (2000) Monophyly and relationships of the Tabanomorpha (Diptera: Bracycera) based on 28S ribosomal gene sequences. *Annals of the Entomological Society of America*, **93**, 1031–1038.

Wiegmann, B.M., Yeates, D.K., Thorne, J.L. & Kishino, H. (2003) Time flies, a new molecular time scale for brachyceran fly evolution without a clock. *Systematic Biology*, **52**, 745–756.

Woodley, N.E. (1989) Phylogeny and classification of the 'Orthorraphous' Bracycera. *Manual of Nearctic Diptera*, Vol. 3 (ed. by J. F. McAlpine & D. V. Wood), pp. 1371–1395. Research Branch, Agriculture Canada, Ottawa.

Yeates, D.K. (1994) Cladistics and classification of the Bombyliidae (Diptera: Asiloidae). *Bulletin of the American Museum of Natural History*, **219**, 1–191.

Yeates, D.K. (2002) Relationships of extant lower Brachycera (Diptera): a quantitative synthesis of morphological characters. *Zoologica Scripta*, **31**, 105–121.

Yeates, D.K. & Wiegmann, B.M. (1999) Congruence and controversy: towards a higher-level phylogeny of Diptera. *Annual Review of Entomology*, **44**, 397–428.

Yeates, D.K. & Wiegmann, B.M. (2005) Phylogeny and evolution of Diptera: recent insights and new perspectives. *The Evolutionary Biology of Flies* (ed. by D. K. Yeates & B. M. Wiegmann), pp. 14–44. Columbia University Press, New York.

Yukuhiro, K., Sezutsu, H., Itoh, H., Shimizu, K. & Banno, Y. (2002) Significant levels of sequence divergence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkmoth, *Bombyx mandarina*, and its close relative, the domesticated silkmoth, *Bombyx mori. Molecular Biology and Evolution*, **19**, 1385–1389.

Accepted 1 March 2006
First published online 15 September 2006