Systematic Entomology

Systematic Entomology (2017), DOI: 10.1111/syen.12275

Taxon sampling to address an ancient rapid radiation: a supermatrix phylogeny of early brachyceran flies (Diptera)

SEUNGGWAN SHIN[1,2], KEITH M. BAYLESS[1,3], SHAUN L. WINTERTON[4], TORSTEN DIKOW[5], BRYAN D. LESSARD[6], DAVID K. YEATES[6], BRIAN M. WIEGMANN[1] and MICHELLE D. TRAUTWEIN[3]

[1]Department of Entomology & Plant Pathology, North Carolina State University, Raleigh, NC, U.S.A., [2]Department of Biological Sciences, University of Memphis, Memphis, TN, U.S.A., [3]California Academy of Sciences, San Francisco, CA, U.S.A., [4]California State Collection of Arthropods, California Department of Food and Agriculture, Sacramento, CA, U.S.A., [5]Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC, U.S.A. and [6]Australian National Insect Collection, CSIRO National Research Collections Australia, Canberra, Australia

Abstract. Early diverging brachyceran fly lineages underwent a rapid radiation approximately 180 Ma, coincident in part with the origin of flowering plants. This region of the fly tree includes 25 000 described extant species with diverse ecological roles such as blood-feeding (haematophagy), parasitoidism, predation, pollination and wood-feeding (xylophagy). Early diverging brachyceran lineages were once considered a monophyletic group of families called Orthorrhapha, based on the shared character of a longitudinal break in the pupal skin made during the emergence of the adult. Yet other morphological and molecular evidence generally supports a paraphyletic arrangement of 'Orthorrhapha', with strong support for one orthorrhaphan lineage – dance flies and relatives – as the closest relative to all higher flies (Cyclorrhapha), together called Eremoneura. In order to establish a comprehensive estimate of the relationships among orthorrhaphan lineages using a thorough sample of publicly available data, we compiled and analysed a dataset including 1217 taxa representing major lineages and 20 molecular markers. Our analyses suggest that 'Orthorrhapha' excluding Eremoneura is not monophyletic; instead, we recover two main lineages of early brachyceran flies: Homeodactyla and Heterodactyla. Homeodactyla includes Nemestrinoidea (uniting two parasitic families Acroceridae + Nemestrinidae) as the closest relatives to the large SXT clade, comprising Stratiomyomorpha, Xylophagidae and Tabanomorpha. Heterodactyla includes Bombyliidae with a monophyletic Asiloidea (exclusive of Bombyliidae) as the closest relatives to Eremoneura. Reducing missing data, modifying the distribution of genes across taxa, and, in particular, removing rogue taxa significantly improved tree resolution and statistical support. Although our analyses rely on dense taxonomic sampling and substantial gene coverage, our results pinpoint the limited resolving power of Sanger sequencing-era molecular phylogenetic datasets with respect to ancient, hyperdiverse radiations.

Correspondence: Seunggwan Shin, Department of Entomology & Plant Pathology, North Carolina State University, Raleigh, NC 27695-7613, U.S.A. E-mail: sciaridae1@gmail.com

Introduction

With revolutionary advances in bioinformatics and molecular genetics, ever larger phylogenomic datasets have become possible, and, in fact, necessary, for resolving major evolutionary questions across the tree of life (Smith *et al.*, 2009; Trautwein *et al.*, 2012; Misof *et al.*, 2014; Yeates *et al.*, 2016; Chesters, 2016). The supermatrix approach, a compilation of all available sequence data into a new, large, single analysis, can be a powerful method to build large-scale phylogenetic trees (de Queiroz & Gatesy, 2007). Because they are opportunistically compiled, supermatrices are often incomplete; in some cases, as much as 70–95% of the data are coded as missing (McMahon & Sanderson, 2006; Smith *et al.*, 2009). Despite this incompleteness, the potential pitfalls of inconsistent data distribution and data overlap (see Dell'Ampio *et al.*, 2014) and the ensuing methodological difficulties, supermatrix methods allow for simultaneous analysis of data from a diversity of sources. Supermatrices have been used to reconstruct phylogenetic relationships across diverse lineages in the tree of life, including kingdoms of life (Ciccarelli *et al.*, 2006), mammals (Meredith *et al.*, 2011), reptiles (Thomson & Shaffer, 2010), birds (Burleigh *et al.*, 2015; Jønsson *et al.*, 2016), various groups of plants (Pirie *et al.*, 2008; Soltis *et al.*, 2011; Hinchliff & Roalson, 2013) and diverse insect clades (van der Linde *et al.*, 2010; Peters *et al.*, 2011; Hedtke *et al.*, 2013; Bocak *et al.*, 2014; Kergoat *et al.*, 2014; Piwczyński *et al.*, 2014). Here we apply supermatrix methods to further resolve the phylogeny of flies by taking advantage of existing molecular data.

The fly tree of life has been addressed at multiple scales of taxonomic and genomic coverage (Wiegmann *et al.*, 2003; Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011; Young *et al.*, 2016). Current estimates of higher-level fly phylogeny reveal three bursts of diversification, with one of these occurring with the emergence of Brachycera approximately 180 Ma, corresponding to the origin of flowering plants (Wiegmann *et al.*, 2011). Most of the brachyceran diversity is included in the large, relatively recent (65 Ma) radiation of Schizophora. Early diverging orthorrhaphous brachyceran lineages include well over 25 000 described species (Pape *et al.*, 2011), and 24 families including familiar groups such as horse flies (Tabanidae), soldier flies (Stratiomyidae), bee flies (Bombyliidae) and robber flies (Asilidae) (Table 1). The rich fossil history of early diverging brachyceran flies begins in the Jurassic (Grimaldi & Cumming, 1999; Grimaldi, 2016), and these lineages include the most species-rich family of blood-feeding insects (Tabanidae), the most species-rich family of predatory flies (Asilidae: Dikow, 2009a,b), parasitoidism (Bombyliidae: bee flies, Acroceridae: small-headed flies, Nemestrinidae: tangle-veined flies), as well as several families of long-tongued ecologically specialized pollinators (Tabanidae, Nemestrinidae, Acroceridae, Bombyliidae: Johnson & Morita, 2006; Karolyi *et al.*, 2012).

The evolution of Brachycera brought about fly lineages with stout bodies and strong aerobatic skills – key characteristics of flies for many people – in contrast to earlier diverging mosquito-like or 'nematocerous' lineages. Brachycera and their sister Bibionomorpha also represent an evolutionary shift from a strong dependence on aquatic environments to terrestriality. The majority of larvae of the earliest 'nematocerous' fly lineages live in aquatic or inundated environments and are saprophagous or phytophagous, whereas the larvae of early brachyceran families are largely land-dwellers that have undergone a major transition to predatory diets, in soil or rotting wood. Larvae of Vermileonidae are notable for specialized predatory behaviour in constructing pit traps. Furthermore, Acroceridae, Nemestrinidae and Bombyliidae larvae are parasitoids of other arthropods. Among parasitoid lineages, there are species that exhibit hypermetamorphosis (at least one larval stage differs from the rest, generally mobile while finding a host, then immobile afterwards) and hyperparasitism (where a parasitoid lives in another parasite). Another distinguishing feature of early brachyceran evolution is less mobile pupae with a more heavily reinforced pupal cuticle that includes strong, posteriorly directed spines and setae that aid in emergence from restrictive substrates such as

Table 1. Previous classification schemes for higher-level groupings of Brachycera, with emphasis on the infraorders and superfamilies of the non-cyclorrhaphan Brachycera based on our results. For clarification of Hilarimorphidae and Apystomyiidae see text.

Family	Infraorder/Superfamily	1[a]	2[a]	3	4[a]	5[a]	6	7	8
Pantophthalmidae	Stratiomyomorpha								
Xylomyidae	Stratiomyomorpha								
Stratiomyidae	Stratiomyomorpha								
Xylophagidae	Xylophagomorpha								
Athericidae	Tabanomorpha								
Austroleptidae	Tabanomorpha								
Bolbomyiidae	Tabanomorpha								
Oreoleptidae	Tabanomorpha								
Pelecorhynchidae	Tabanomorpha								
Rhagionidae	Tabanomorpha								
Tabanidae	Tabanomorpha								
Vermileonidae	Tabanomorpha								
Acroceridae	Nemestrinoidea								
Nemestrinidae	Nemestrinoidea								
Apioceridae	Asiloidea								
Apsilocephalidae	Asiloidea								
Asilidae	Asiloidea								
Evocoidae	Asiloidea								
Mydidae	Asiloidea								
Scenopinidae	Asiloidea								
Therevidae	Asiloidea								
Bombyliidae	Asiloidea?								
Hilarimorphidae	Unplaced								
Apystomyiidae	Unplaced								
Empididae	Empidoidea								
Atelestidae	Empidoidea								
Dolichopodidae	Empidoidea								
Hybotidae	Empidoidea								
Cyclorrhapha	Cyclorrhapha								

[a]Stable positions in our analyses (Excluding unplaced Hilarimorphidae and Apystomyiidae).

1, SXT clade; 2, Homeodactyla; 3, Muscomorpha sensu Woodley, 1989; 4, Heterodactyla; 5, Eremoneura; 6, Platygenya (Orthorrhapha sensu Wiegmann *et al.* (2011)); 7, Orthopyga sensu Aczél, 1954; 8, Orthorrhapha (non-cyclorrhaphan brachyceran families).

 doi: 10.1111/syen.12275

soil and decaying wood; notwithstanding this, the pupae of early brachyceran lineages are still capable of movement. This contrasts with the immobile and unadorned pupae of cyclorrhaphan flies that are encased in the hardened and capsule-like last larval cuticle – the puparium (Woodley *et al.*, 2009).

The phylogenetic relationships of early diverging brachyceran fly families have been a challenge to resolve well into the era of molecular phylogenetics (Yeates & Wiegmann, 1999; Wiegmann *et al.*, 2003; Wiegmann *et al.*, 2011). Non-cyclorrhaphan brachyceran families (Table 1) have traditionally been grouped together as Orthorrhapha. The name Orthorrhapha derives from the straight line of breakage in the pupal skin made during the emergence of the adult, in contrast to the circular exit hole in the puparium that characterizes Cyclorrhapha (Brauer, 1880; Lameere, 1906). Despite this characteristic, Orthorrhapha has long been suspected to be paraphyletic, lacking both robust morphological and molecular support (Woodley, 1989; Yeates and Wiegmann, 1999; Yeates, 2002; Wiegmann *et al.*, 2003); in fact, a monophyletic Orthorrhapha in the traditional sense (all non-cyclorrhaphan brachycerans, including Empidoidea) has not been advanced by any recent study.

In contrast to the concept of a monophyletic Orthorrhapha, morphological and molecular studies generally show evidence in favour of joining the orthorrhaphan infraorder Empidoidea (dance flies and relatives) as the sister group to Cyclorrhapha (e.g. Woodley, 1989), together called Eremoneura. In addition to Eremoneura, there is morphological and molecular evidence for Heterodactyla, a group uniting Asiloidea (robber flies and relatives; see Table 1) and Eremoneura (Woodley, 1989). The name Heterodactyla refers to all brachyceran flies with the empodium, the medial tarsal lobe, reduced to be bristle-like or lost (Lambkin *et al.*, 2013). The division of Brachycera based on a bristle-like empodium led to a reciprocal uniting of all groups with a pulvilliform empodium (specifically narrow mediolobus) under the term Homeodactyla (Stuckenberg, 2001) – Stratiomyomorpha, Tabanomorpha, Nemestrinoidea, and Xylophagidae – yet this designation was made without any consideration as to its monophyly. Contrary to Homeodactyla is the more widely accepted Muscomorpha – a clade composed of Heterodactyla + Nemestrinoidea (Nemestrinidae + Acroceridae) (Woodley, 1989) (Table 1), that is supported by multiple apparent morphological synapomorphies, including the antennal flagellum reduced to four or fewer articles, loss of tibial spurs, female cerci 1-segmented, and the structure of the larval antennae (Lambkin *et al.*, 2013).

The first large-scale molecular analysis to address higher-level fly phylogeny returned unexpected results regarding non-cyclorrhaphan brachycerans, with support for a monophyletic Orthorrhapha excluding Empidoidea, as the sister group to Eremoneura (Empidoidea + Cyclorrhapha) (Wiegmann *et al.*, 2011). Orthorrhapha excluding Empidoidea were termed Platygenya by Brauer (1883, Table 1), but this term was not widely adopted, so throughout we define Orthorrhapha as Orthorrhapha sensu Wiegmann *et al.* (2011), excluding Empidoidea. In Wiegmann *et al.* (2011), a monophyletic Orthorrhapha was supported only in analyses that maximized taxon sampling, whereas analyses of more genes but fewer taxa (Wiegmann *et al.*, 2011: Figure S1), or analyses of morphological data alone (Lambkin *et al.*, 2013) recovered a paraphyletic arrangement of the early brachyceran lineages.

Aside from the monophyly of Orthorrhapha, questions about the relationships between major families of non-cyclorrhaphan Brachycera (reviewed in Woodley, 1989; Yeates, 2002; Wiegmann *et al.*, 2003, 2011; Sinclair *et al.*, 2013) (Fig. 1) remain to be resolved. Not settled yet, for instance, is the question of whether the two parasitic families Nemestrinidae (parasitoids of insects) and Acroceridae (parasitoids of spiders) are a monophyletic group (= Nemestrinoidea, Woodley, 1989). Weak morphological evidence supports Nemestrinoidea (Woodley, 1989), whereas molecular analyses place the three parasitic families (including Bombyliidae) as distantly related (Winterton *et al.*, 2007; Wiegmann *et al.*, 2011). Another questionable higher-level grouping is the relationship between three infraorders Stratiomyomorpha (soldier flies and relatives), Xylophagomorpha (only including Xylophagidae) and Tabanomorpha; these are sometimes gathered together in what is known as the SXT clade (Yeates, 2002; Wiegmann *et al.*, 2003), although this clade is often weakly supported by both molecules and morphology. The monophyly of Asiloidea including Bombyliidae is another outstanding hypothesis that has weak support based on morphology and has been challenged by molecular data (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011).

Other key questions regarding early brachyceran phylogeny involve the placement of two unusual and rarely-collected fly genera – *Apystomyia* Melander and *Hilarimorpha* Schiner. On the one hand, *Apystomyia* was formerly considered a putative member of Asiloidea and grouped in Bombyliidae or together with *Hilarimorpha* in the family Hilarimorphidae, but has recently and consistently been placed as the closest relative to the large radiation of Cyclorrhapha, based on molecular data (Trautwein *et al.*, 2010, Wiegmann *et al.*, 2011). Morphological interpretations of the male genitalia of *Apystomyia* place it instead as the sister group to Eremoneura (Empidoidea + Cyclorrhapha; see Yeates, 2002; Sinclair *et al.*, 2013). *Hilarimorpha*, on the other hand, remains completely uncertain in its phylogenetic placement, with multiple ambiguous alternatives found in nearly all molecular (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011) and morphological (Yeates, 1994, 2002) analyses.

The aim of this study is to recover the phylogenetic relationships of early diverging brachyceran fly lineages by compiling and analysing existing nucleotide data to determine whether increased taxon sampling (even in a patchy supermatrix) can confirm or refute existing phylogenetic hypotheses. We constructed a supermatrix using data from previous studies (see Materials and Methods) and added taxa to the ingroup and outgroup using methods for supermatrix data mining from GenBank (outlined in Peters *et al.*, 2011). We evaluated the impact of minimizing missing data, differential taxon coverage, and excluding rogue taxa in various analyses. Our findings are based on the largest molecular dataset that has been compiled to date for resolving the phylogenetic relationships of flies.

 doi: 10.1111/syen.12275

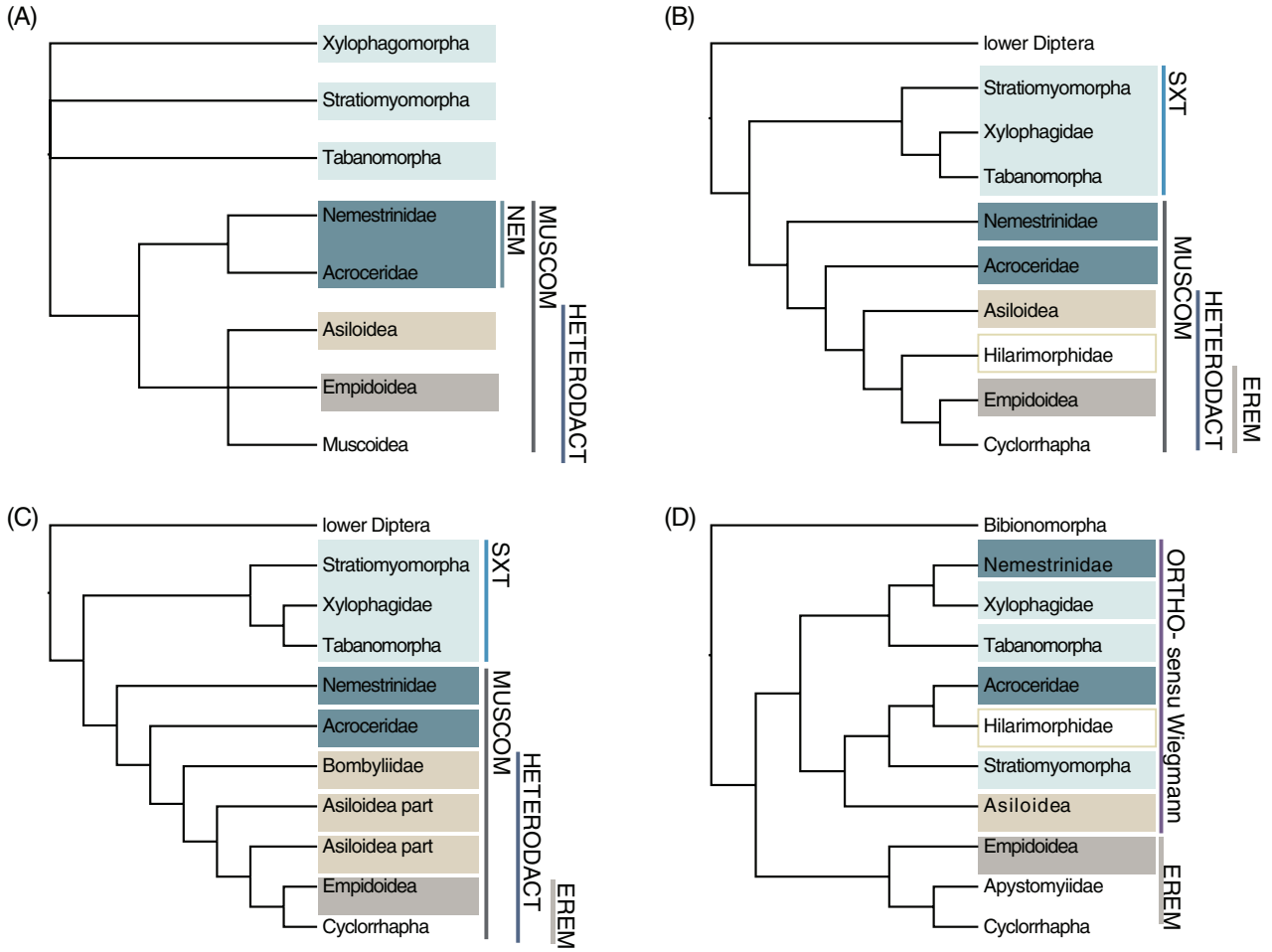

Fig. 1. History of Brachycera phylogeny focusing on non-cyclorrhaphan brachyceran flies. (A) Woodley's (1989) morphology; (B) Yeates' (2002) morphology; (C) Wiegmann *et al.*'s (2003) morphology + molecular (part 1: Asilidae, Apioceridae, Mydidae; Part 2: Therevidae, Scenopinidae); (D) Wiegmann *et al.* 's (2011) morphology + molecular. SXT, Stratiomyomorpha + Xylophagidae + Tabanomorpha; EREM, Eremoneura; NEM, Nemestrinoidea (Acroceridae + Nemestrinidae); MUSCOM, Muscomorpha; HETERODACT, Heterodactyla; ORTHO-sensu Wiegmann, Orthorrhapha sensu Wiegmann *et al.*, 2011. [Colour figure can be viewed at wileyonlinelibrary.com].

Materials and methods

Sampling design

We included a total of 1217 species of outgroup and ingroup taxa in our supermatrix analyses (Table S1). Our dataset included 399 species of Asiloidea, 102 species of Stratiomyomorpha, 224 species of Tabanomorpha, 43 species of Acroceridae, seven species of Nemestrinidae, eight species of Xylophagidae, one species of Apystomyiidae and one species of Hilarimorphidae from previous phylogenetic studies. We additionally downloaded and parsed sequences up to May 2014 using scripts published in Peters *et al.* (2011) to include further ingroup species of Eremoneura (368 Empidoidea and 50 Cyclorrhapha) and included in total 14 species belonging to Bibionomorpha, representing nine distinct families of Neodiptera (uniting Brachycera + Bibionomorpha, see Wiegmann *et al.*, 2011) as outgroup taxa (Figure S7).

Alignments

We selected 20 genes previously used for phylogenetic estimates addressing non-cyclorrhaphan brachyceran fly lineages that are well represented in GenBank. These sequences include 14 nuclear protein-coding genes (*AATS1*, *AATS2*, *ACE-1*, *CAD*, *EF1A*, *G6PD*, *PEPCK*, *PER*, *PGD*, *PUG*, *Rhodopsin*, *SINA*, *SNF* and *TPI*), the small and large subunit of nuclear ribosomal RNA genes *18S* rRNA and *28S* rRNA, two mitochondrial protein-coding genes (*COI* and *COII*), and the small and large subunits of the mitochondrial ribosomal RNA genes *12S* rRNA and *16S* rRNA. Sequences of previously published datasets were collected from the following studies: Yang *et al.*, 2000; Winterton *et al.*, 2001, 2007; Wiegmann *et al.*, 2003; Brammer & von Dohlen, 2007; Dikow, 2009a; Trautwein *et al.*, 2010, 2011; Wiegmann *et al.*, 2011; Lessard *et al.*, 2013; Winterton & Ware, 2015; Morita *et al.*, 2016. All sequences for each study were re-aligned using MAFFT (Katoh &

 doi: 10.1111/syen.12275

Table 2. Descriptive statistics for alignments and trees resulting from each of six combined approaches to data filtering.

Dataset	Methods	#Species	#Genes	Alignment length (nt, 1 + 2 codon position)	Missing (%)	Decisiveness
(i)	All taxa all genes; rogues included	1217	20	21 369	83	0.80
(ii)	More than 40 taxa per gene; rogues included	1217	10	17 516	68	0.80
(iii)	More than three infraorders per gene; rogues included	1208	14	18 998	80	0.81
(iv)	All taxa all genes; rogues excluded	1095	20	21 369	82	0.82
(v)	More than 40 taxa per gene; rogues excluded	1095	10	17 516	66	0.81
(vi)	More than three infraorders per gene; rogues excluded	1087	14	18 998	79	0.83

(i) ALL; (ii) MISSING; (iii) COVERAGE; (iv) ALL_ROGUE; (v) MISSING_ROGUE; (vi) COVERAGE_ROGUE.

Standley, 2013) with the Q-INS-i algorithm for the mitochondrial and nuclear rRNA genes, and with the FFT-NS-i algorithm for all protein-coding genes. Alignments were checked by eye using MEGA 6.06 (Tamura *et al.*, 2013) and SEAVIEW 4 (Gouy *et al.*, 2010). Obvious errors or frame shifts were corrected manually. We subsequently added sequences from ingroup species belonging to Eremoneura and outgroup species belonging to Bibionomorpha using the profile alignment strategy 'MAFFT --add' (Katoh & Standley, 2013) individually for each gene via the web server (v7; http://mafft.cbrc.jp/alignment/server/add_sequences.html) with L-INS-1 for rRNA with settings enabled to maintain secondary structure inferences for the previously curated rRNA datasets and using FFT-NS-1 for protein-coding genes. The third codon positions were removed from protein coding genes using the Perl script 'selectSites.pl' (http://raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/perl-scripts.html) to reduce the high level of noise included in these sites (see Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011 for third position heterogeneity). SEQUENCEMATRIX OSX 1.7.8 (Vaidya *et al.*, 2011) was used to concatenate all genes. Duplicates were checked by eye in a text editor and duplicate sequences/taxa were removed based on high sequence identity over a defined sequence length in SEQUENCEMATRIX OSX 1.7.8 (Vaidya *et al.*, 2011). We used only the longest sequence for each gene available for a particular taxon. Our largest supermatrix included 1217 taxa and 20 genes comprising an alignment length of 21 369 sites of which 83% was coded as missing (Table 2).

Data filtering and tree searching

We analysed six datasets to evaluate the robustness of inferred phylogenies under different treatments to address missing data, taxonomic coverage of genes and rogue taxa removal. Our datasets are as follows (Table 2):

(i) Dataset '**ALL**' includes all compiled data – 1217 species; 20 genes; spanning an alignment length of 21 369 nt sites (rRNA genes. *12S*: 410, *16S*: 1552, *18S*: 4117, *28S*: 5984; first and second codon positions. *AATS1*: 338 bp, *AATS2*: 1158 bp, *ACE-1*: 74 bp, *COI*: 964, *COII*: 242, *CAD*: 2580 bp, *EF1A*: 800 bp, *G6PD*: 480 bp, *PEPCK*: 242 bp, *PER*: 460 bp, *PGD*: 458 bp, *PUG*: 382 bp, *Rhodopsin*: 315 bp, *SINA*: 278 bp, *SNF*: 222 bp and *TPI*: 313 bp); and 83% coded as missing data.

(ii) Dataset '**MISSING**' includes the ten genes with the largest fraction of present data to minimize the amount of data coded as missing. The dataset comprises 1217 species; ten genes (*12S*, *16S*, *18S*, *28S*, *AATS1*, *CAD*, *COI*, *EF1A*, *PGD* and *TPI*); 17 516 nt sites (RNA genes, first and second codon positions); 68% missing data.

(iii) Dataset '**COVERAGE**' includes only genes that were sampled from three or more orthorrhaphan infraorder taxa: 1208 species; 14 genes (*16S*, *18S*, *28S*, *AATS1*, *AATS2*, *CAD*, *COI*, *ACE-1*, *G6PD*, *PEPCK*, *PER*, *PGD*, *SINA* and *TPI*); 18 998 nt sites (RNA genes, first and second codon positions); 80% missing data.

From each dataset we constructed an additional subsequent dataset by removing the top 10% of instability scored taxa (or rogue taxa) (Table S1). We measured leaf stability of included species using 'instability_multicore.py' (http://dx.doi.org/10.5061/dryad.6p76c3pb) to rapidly calculate instability scores for large sets of very large trees (Hinchliff & Roalson, 2013). Our datasets with rogue taxa removed are as follows (Table 2):

(iv) Dataset '**ALL_ROGUE**' includes all compiled data containing 1095 species after excluding rogue taxa; 20 genes; spanning an alignment length of 21 369 nt sites (RNA genes, first and second codon positions); with 82% missing data.

(v) Dataset '**MISSING_ROGUE**' includes the ten genes with the largest fraction of present data to minimize the amount of data coded as missing. The dataset comprises 1095 species after excluding rogue taxa; ten genes (*12S*, *16S*, *18S*, *28S*, *AATS1*, *CAD*, *COI*, *EF1A*, *PGD* and *TPI*); 17 516 nt sites (RNA genes, first and second codon positions); 66% missing data.

(vi) Dataset '**COVERAGE_ROGUE**' includes only genes that were sampled from three or more orthorrhaphan infraorder taxa: 1087 species; 14 genes (*16S*, *18S*, *28S*, *AATS1*, *AATS2*, *CAD*, *COI*, *ACE-1*, *G6PD*, *PEPCK*, *PER*, *PGD*, *SINA* and *TPI*); 18 998 nt sites (RNA genes, first and second codon positions); 79% missing data.

RAxML v8.2.9 (Stamatakis, 2014) was employed to infer maximum-likelihood (ML) trees from each concatenated

 doi: 10.1111/syen.12275

supermatrix dataset (i–vi) on the North Carolina State University Bioinformatics Research Center cluster (NCSU BRC cluster http://brccluster.statgen.ncsu.edu/). First, we performed ten single ML tree searches with a random start tree, with the dataset partitioned by gene, with linked branch lengths to find the best likelihood tree for each dataset. We tested both approaches, GAMMA and CAT with the substitution model GTR on the largest dataset (ALL) and using only GTR + GAMMA (with default four rate categories) for remaining datasets because parameter estimation is ostensibly more accurate. Next, we applied 100 thorough nonparametric bootstrap replicates to estimate branch support with the GTR + GAMMA model. After analysis, we also performed 1000 partitioned thorough nonparametric bootstrap replicates for a representative tree from dataset MISSING_ROGUE (v) (Fig. 3, Figure S5, tree was selected by topology assessments of majority rule consensus trees in Fig. 2). The bootstrap convergence was tested (-I autoMRE option by RAxML) with 1000 BT (trees converged after 300 bootstraps). Resulting bootstrap support was mapped onto the best ML trees (Figures S1–S6). Finally, partial tree-wise decisiveness (PDC) was calculated using the program 'decisivator' (https://github.com/josephwb/Decisivator) for our six data sets. To compare branches showing >50% ML Bootstrap Support (MLBS) from each tree, we assessed topologies with majority rule consensus trees from the 100 bootstrap replicates for each dataset using RAxML (Fig. 2). ML trees and concatenated Phylip files are provided as Supplementary materials (Dryad accession #; https://doi.org/10.5061/dryad.cq63m).

Results and Discussion

Here we present the most densely sampled phylogeny to date, in terms of species (1217) and genes (20) addressing relationships of non-cyclorrhaphan Brachycera. Our topologies from all six ML analyses recover strikingly similar arrangements of families across non-cyclorrhaphan brachyceran flies. A monophyletic Orthorrhapha is never recovered (see Wiegmann *et al.*, 2011) [Fig. 2, Figures S1–S7. S1: (i) ALL; S2: (ii) MISSING; S3: (iii) COVERAGE; S4: (iv) ALL_ROGUE; S5: (v) MISSING_ROGUE; S6: (vi) COVERAGE_ROGUE; see Materials and methods for details on each dataset].

Although the topologies that we recovered are consistent across all datasets, the bootstrap support for deeper relationships in the brachyceran tree is weak, despite extensive taxon sampling. We find that reducing the amount of missing data and therefore increasing data overlap (either by total percentage, or by maximizing the inclusion of genes that have broad taxonomic coverage) has mixed effects on bootstrap support for major clades. In general, an increase in data overlap, in terms of an increasing amount of data present, does not substantially increase node support (Table 3). In contrast, removing rogue taxa alone, or removing rogues along with reducing the amount of missing data, appears to substantially increase branch support across the tree, as well as decisiveness (for further discussion on the term 'decisiveness' in the context of phylogenomic analyses, see Steel & Sanderson, 2010; Sanderson *et al.*, 2010) (Table 2). Note that we did not address the distribution of missing data that also can have a major impact if distributed unevenly (see Misof *et al.*, 2013).

Our findings both confirm and contest previous hypotheses of non-cyclorrhaphan brachyceran relationships based on morphological and molecular data. We consistently recover two primary brachyceran lineages: Homeodactyla and Heterodactyla (Table 1). Homeodactyla includes a monophyletic SXT clade (including monophyletic Stratiomyomorpha, monophyletic Xylophagidae and monophyletic Tabanomorpha) as the closest relative to a monophyletic parasitoid clade, Nemestrinoidea (Nemestrinidae + Acroceridae) (Figures S1–S6). Within Heterodactyla, we consistently recover a monophyletic Eremoneura and a monophyletic Asiloidea, excluding monophyletic Bombyliidae [datasets (i) and (v) only; *Evocoa chilensis* is spuriously sister taxon to Eremoneura in all other analyses]. Bombyliidae, a notoriously rogue lineage, evades stable placement, yet is always placed close to either Asiloidea or Eremoneura, congruent with earlier studies (e.g. Wiegmann *et al.*, 2003, 2011). Also, we confirm a sister-group relationship between the small, rare family Apystomyiidae and Cyclorrhapha. Lastly, the enigmatic family Hilarimorphidae remains somewhat phylogenetically ambiguous as it is placed as the closest relative to Homeodactyla across all analyses but with very low branch support (Table 3, Figures S1–S6).

Homeodactyla

Homeodactyla, uniting the SXT clade + Nemestrinoidea, is recovered in all of our analyses but with poor bootstrap support (ranging from 25 to 48%). This clade unites brachyceran flies with pulvilliform empodium, a character state in which the medial lobe on the pretarsus is pad-like. Yet, Homeodactyla contradicts the morphology-based infraorder Muscomorpha sensu Woodley (1989) (i.e. Nemestrinoidea (Asiloidea, Eremoneura)) that is supported in part by multiple adult characters including male genitalic (Woodley, 1989; Yeates & Wiegmann, 1999).

SXT clade. All of our analyses recover a monophyletic SXT clade joining Stratiomyomorpha, Xylophagidae and Tabanomorpha, although the highest bootstrap support is 48% (Table 3). Previous molecular analyses showed weak or no support for an SXT clade (Wiegmann *et al.*, 2003, 2011), and likewise strong morphological synapomorphies for the group are lacking (Woodley, 1989; Yeates & Wiegmann, 1999).

The monophyly and placement of infraorder Stratiomyomorpha found by our analyses corresponds with the morphological hypothesis of Yeates (2002) and molecular analyses of Wiegmann *et al.* (2003) (Fig. 1). An artefactual placement of the species *Solva marginata* (Meigen) (Xylomyidae) within Stratiomyidae is likely due to the limited data available for this species. *Solva marginata* is represented only by *12S* and *16S* mitochondrial ribosomal RNA; another representative *Solva* sp. CB0434 (*28S* nuclear ribosomal RNA and nuclear protein coding *EF1A*) grouped with Xylomyidae as the sister of *Xylomya*

 doi: 10.1111/syen.12275

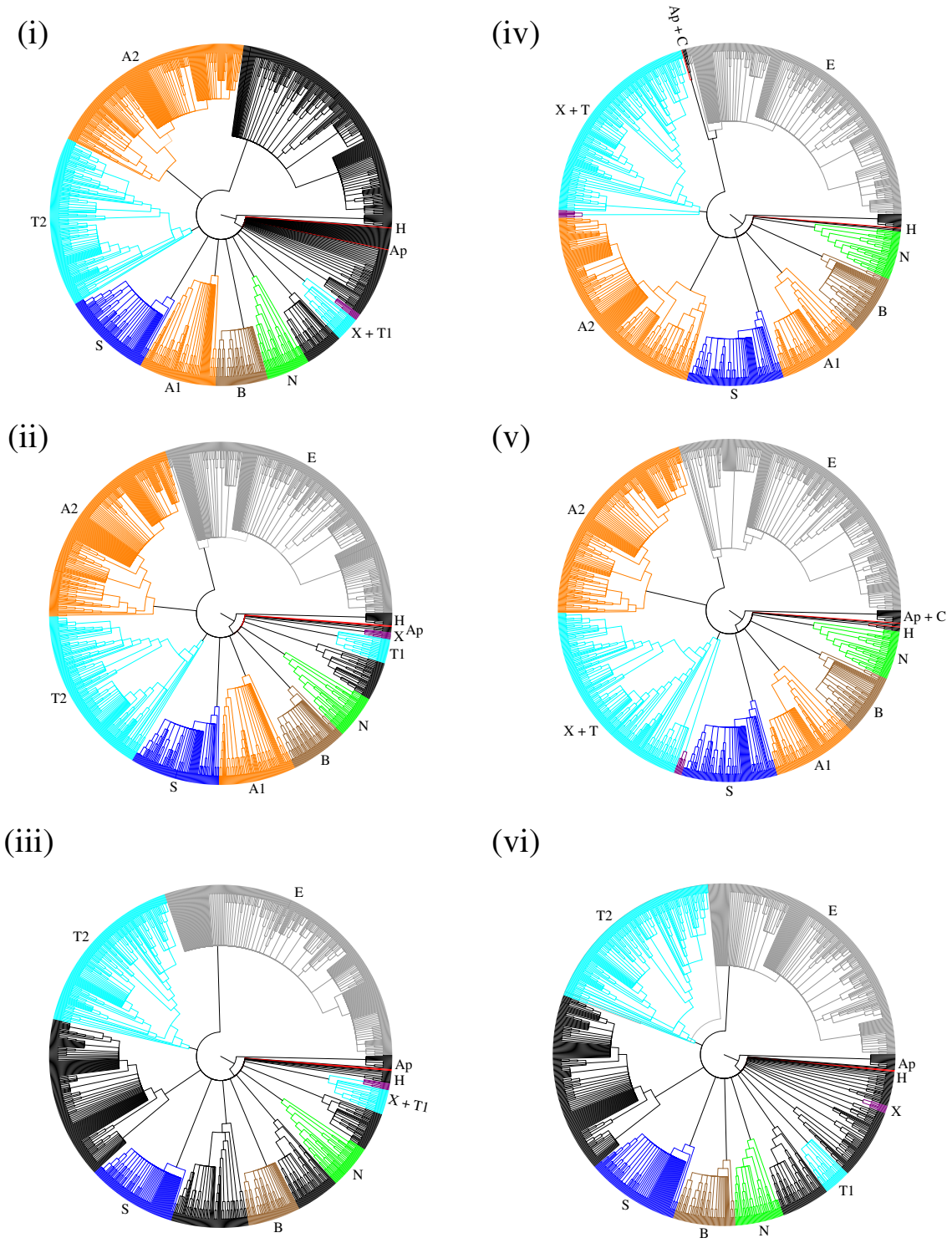

Fig. 2. Fifty percent majority-rule consensus trees from 100 ML thorough bootstrap replicates performed for each of the supermatrices in Table 2. Trees and datasets: (i) ALL; (ii) MISSING; (iii) COVERAGE; (iv) ALL_ROGUE; (v) MISSING_ROGUE; (vi) COVERAGE_ROGUE in Table 2. Ap, *Apystomyia*; B, Bombyliidae; C, Cyclorrhapha; E, Empidoidea; H, *Hilarimorpha*; N, Nemestrinoidea (Acroceridae + Nemestrinidae); A1, Asilidae + Mydidae; A2, Scenopinidae + Therevidae; T1, Austroleptidae + Bolbomyiidae + Rhagionidae + Vermileonidae; T2, Athericidae + Oreoleptidae + Pelecorhynchidae + Tabanidae. Comparing the topology of (i) ('ALL') to each successive topology, resolution is increased by the reduction of missing data and the removal of rogue taxa. [Colour figure can be viewed at wileyonlinelibrary.com].

 doi: 10.1111/syen.12275

Table 3. Clade recovery based on MLBS (%) in analyses including each of three combined approaches with pre (i, ii, iii) and post (iv, v, vi) removal of rogue analyses.

Dataset	Br	Pl	Em	Er	Ap + Cy	As	Ta	Xy	St	SXT	Ne	He	Ho	Bo	(Bo(As(Er, Ap)))	(Hi, (Ho))
(i)	96	-	41	9	27	20	29	98	86	48	66	36	37	47	36	17
(ii)	98	-	17	8	35	15	33	95	89	48	69	34	38	19	2	18
(iii)	74	-	67	18	25	15	23	100	76	41	59	28	25	40	28	12
(iv)	89	-	87	52	57	27	42	100	96	47	76	22	48	57	22	16
(v)	94	-	71	41	55	38	48	100	94	42	70	29	42	63	29	13
(vi)	53	-	66	11	17	15	28	99	85	40	59	3	36	63	31	19

(i) ALL; (ii) MISSING; (iii) COVERAGE; (iv) ALL_ROGUE; (v) MISSING_ROGUE; (vi) COVERAGE_ROGUE; Br, Brachycera; Pl, Platygenya; Em, Empididae; Er, Eremoneura; Ap, *Apystomyia*; Cy, Cyclorrhapha; As, Asiloidea [underlined = nonmonophyletic Asiloidea; *Evocoa chilensis* is sister to Er]; St, Stratiomyomorpha; Xy, Xylophagidae; Ta, Tabanomorpha; SXT, Stratiomyomorpha + Xylophagidae + Tabanomorpha; Ne, Nemestrinoidea; Bo, Bombyliidae; He, Heterodactyla; Hi, Hilarimorphidae; Ho, Homeodactyla.

Rondani (Figure S5), and this position is more plausible (Brammer & von Dohlen, 2010).

The relationship of Tabanomorpha and Xylophagidae is largely congruent with recent morphological and molecular phylogenetic results from Yeates (2002) and Wiegmann *et al.* (2003). Rhagionidae are paraphyletic and the superfamily Rhagionoidea may include Vermileonidae, Austroleptidae and Bolbomyiidae.

Nemestrinoidea. A consistent result is the recovery of a monophyletic, moderately supported Nemestrinoidea, with Nemestrinidae and Acroceridae as sister groups (bootstrap support 59–76%, Table 3). This finding is in agreement with the concept of Nemestrinoidea from Woodley (1989) but rejects Hennig (1973) who also included Bombyliidae, which is the other non-cyclorrhaphan brachyceran family of parasitoids (Yeates & Greathead, 1997). Therefore, hypermetamorphic parasitoidism in Nemestrinoidea and Bombyliidae is likely to have evolved independently.

The recovery of a monophyletic Nemestrinoidea is surprising in the sense that this result is in contrast to all previous morphological and molecular phylogenetic analyses that have addressed the relationship between Nemestrinidae and Acroceridae (Yeates, 2002; Wiegmann *et al.*, 2003; Winterton *et al.*, 2007; Wiegmann *et al.*, 2011). Previous findings based on molecular data place Nemestrinidae separate from Acroceridae in various arrangements, with both families paraphyletic with respect to Asiloidea (Winterton *et al.*, 2007), or with Nemestrinidae as the closest relative to Xylophagidae (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011). Also, from a morphological perspective, Woodley (1989) noted that Nemestrinoidea are only weakly supported by the presence of parasitic, hypermetamorphic larvae (shared by Bombyliidae).

Heterodactyla

Heterodactyla, the grouping of Asiloidea (as well as Bombyliidae) and Eremoneura (Empidoidea + Cyclorrhapha), were recovered in all analyses, but with negligible bootstrap support – 36% at its highest (ALL dataset; see Figure S1, Table 3). Heterodactyla, as opposed to the hypothesized Muscomorpha sensu Woodley (1989) (i.e. Nemestrinoidea (Asiloidea, Eremoneura)), unites all brachycerans with a reduced or absent tarsal empodium (notwithstanding the challenging Hilarimorphidae). Asiloidea are consistently recovered as a clade, albeit with the exclusion of Bombyliidae (Table 3). This result is congruent with previous molecular analyses that have supported a monophyletic Asiloidea excluding Bombyliidae (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011). Within Asiloidea, Mydidae were found to be paraphyletic, with the genus *Tongamya* Stuckenberg separate from the remaining mydids (Fig. 3, Figures S1–S5, *Tongamya* sister to the remaining Mydidae are supported 0% MLBS in Figure S6). This arrangement is in conflict with previous studies based on *28S* rDNA and morphological data (Irwin & Wiegmann, 2001; Dikow, 2009b, but see Trautwein *et al.*, 2010). *Tongamya* was previously placed within Apioceridae, yet currently is considered to be an early diverging mydid lineage (Yeates & Irwin, 1996). No other early diverging mydid lineages were included in this supermatrix analysis.

A putative splitting of Brachycera into Homeodactyla and Heterodactyla upends hypotheses about 'Orthorrhapha' being a comb of lineages, from the saprophagous Stratiomyomorpha to Tabanomorpha with predatory larvae and many blood-feeding adults, to Asiloidea with predatory or parasitoid larvae and some predatory adults. Instead, our findings imply that two lineages of Brachycera followed independent evolutionary trajectories. Considering the limited bootstrap support, however, this hypothesis requires more investigation.

Progress on the placement of small, phylogenetically ambiguous families

Apystomyiidae, a monotypic family with only one known extant species, *Apystomyia elinguis* Melander, but several extinct species (Grimaldi, 2016), was re-collected in 2005 after decades of evading capture since its original discovery in the 1940s and description by Melander in 1950 (Melander, 1950). All subsequent molecular analyses have found *Apystomyia* to be the sister group to the large radiation of Cyclorrhapha with high support (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011). We also find *Apystomyia* as closest relative to Cyclorrhapha in

 doi: 10.1111/syen.12275

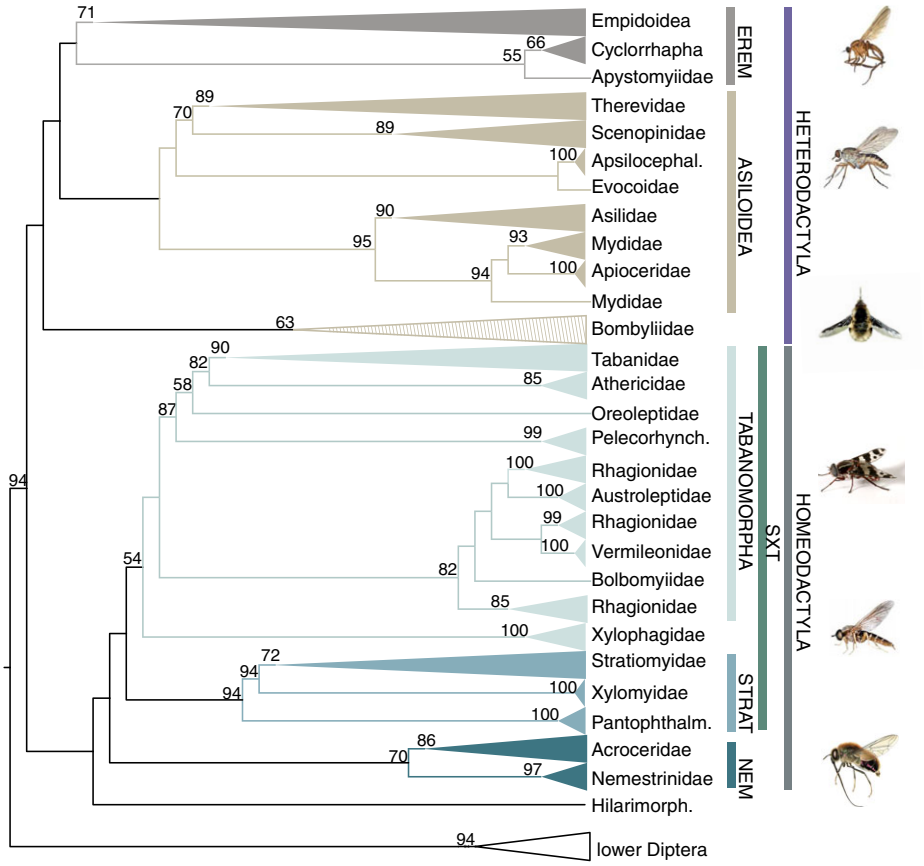

Fig. 3. Representative ML tree with branch support from 100 thorough bootstrap replicates for our best-resolved dataset (v) 'MISSING_ROGUE'. This dataset includes half of the sampled genes, excluding ten genes with the largest fraction of missing data: 1095 taxa, ten genes, 17 516 nt sites (first and second codon positions) with 66% missing data. Bootstrap support >50% is displayed. The original tree with full names of taxa submitted as Figure S5. Collapsed clades are indicated by triangles. SXT, Stratiomyomorpha + Xylophagidae + Tabanomorpha; EREM, Eremoneura; STRAT, Stratiomyomorpha; NEM, Nemestrinoidea (Acroceridae + Nemestrinidae); Apsilocephal., Apsilocephalidae' Pelecorhynch., Pelecorhynchidae; Pantophthalm., Pantophthalmidae; Hilarimorph., Hilarimorphidae. [Colour figure can be viewed at wileyonlinelibrary.com].

the majority of our analyses, although lacking strong bootstrap support. Morphological interpretations of the male genitalia of *Apystomyia* place it instead as sister group to Eremoneura (Yeates, 2002; Sinclair *et al.*, 2013). In two analyses (ALL and COVERAGE_ROGUE Figures S1 and S6, MLBS <17%), the cyclorrhaphan *Lonchoptera* Meigen spuriously grouped with *Apystomyia*. The genus *Lonchoptera* was represented only by a single fragment of *28S* rRNA (Table S1) and all representatives were removed as rogue taxa in further analyses of ALL and MISSING (Table S1). Further phylogenomic analyses are needed to confirm the placement of the rarely-collected *Apystomyia* fly.

In congruence with previous studies, we find that the phylogenetic placement of Hilarimorphidae cannot be elucidated confidently. Hilarimorphidae is the sister group to Homeodactyla in our analysis and had similar placement in Wiegmann *et al.* (2011), but in all cases with negligible bootstrap support (<19%, Table 3). We do not recover (*Hilarimorpha* + *Apystomyia*) as a clade close to Bombyliidae as suggested by Yeates (1994). Although there are some morphological characters that corroborate a relationship between Hilarimorphidae and Eremoneura (abdominal tergite 9 is absent in females of both Hilarimorphidae and Eremoneura, and both lineages lack the lateral ejaculatory processes (Lambkin *et al.*, 2013)), phylogenetic inferences using morphological data have not clearly solved the phylogenetic affinities of Hilarimorphidae (Sinclair & Cumming, 2006; Sinclair *et al.*, 2013). The larvae of Hilarimorphidae, another potential source of phylogenetically informative morphological characters, are unknown. Hilarimorphidae have not been consistently placed with strong support in any molecular phylogeny to date (Trautwein *et al.*, 2010; Wiegmann *et al.*, 2011) and therefore remain the most phylogenetically ambiguous non-cyclorrhaphan brachyceran fly family.

Challenges for supermatrix analyses

Lineage- and gene-specific differences in phylogenetic information, evolutionary rate and undetected paralogy are critical problems for constructing massive, taxon-rich supermatrices.

 doi: 10.1111/syen.12275

We sourced data in two different ways for our supermatrix compilation: (i) we used previously curated datasets from published studies of non-cyclorrhaphan brachyceran phylogeny, and (ii) downloaded additional sequences from NCBI using criteria and methods recommended by Peters *et al.* (2011) in their supermatrix construction pipeline. Sequences for Eremoneura (Empidoidea + Cyclorrhapha) and nematocerous outgroups were obtained through NCBI. Similar to the mega-phylogeny approach (Smith *et al.*, 2009), we chose to include only genes from NCBI that had previously been used in phylogenetic analyses.

The previously curated datasets were carefully compiled by expert systematists in each group. In contrast, the sequences acquired by data mining GenBank were mostly algorithmically trimmed. In compiling our previously curated dataset, MAFFT (Katoh & Standley, 2013) was especially valuable due to its incorporation of secondary structure-considered alignment option for ribosomal genes (Q-INS-i). In an initial set of alignment estimates that treated the fully concatenated dataset as a single alignment, the ribosomal gene sections yielded spurious alignments with large numbers of gaps that would have to be corrected by manual alignment. We then used MAFFT --add, a profile alignment option, which allowed us to constrain previously curated data blocks so that we could subsequently add large numbers of variable length sequences available in NCBI. Alignment complexity due to the availability of highly variable gene fragments may be the most important challenge for the construction of meaningful large datasets from existing data. All applications of automated alignment programs for our datasets resulted in misaligned sequences that required modification by translation and manual editing.

The quality differences in our previously curated data as compared to GenBank-sourced data are evident in our analyses and results. One of our most notable results is that rogue taxa have a strong impact on topology and support values in comparison to the impact of missing data (Fig. 2, Tables 2, 3); as it turns out, the majority (97%) of rogue taxa were mined from GenBank, whereas the rogue taxa from previously curated datasets make up only 3% of all filtered rogues (Table S1). Also, 50% of filtered rogue taxa from dataset (i) are represented by only one gene (Table S1) (see Brower, 2017). Furthermore, the phylogenetic resolution of the GenBank-sourced data is poor, resulting in a polytomy for Eremoneura in all majority-rule consensus trees (Fig. 2). In contrast, the previously curated data for other brachyceran lineages provide a largely resolved tree supported in some cases with moderately high (>50) bootstraps. We suspect that the poor performance of the GenBank-sourced data (e.g. Chesters, 2016) in large part results from poor alignment due to variable sequence quality, length or amplicons originating from different gene fragments, and/or possibly misidentified chimeric species based on different genes.

Missing data and rogue taxa. A high percentage of missing data and rogue taxa can be attributes of supermatrix analyses that pull together data from published studies using Sanger data. The most resolved, well-supported tree in our study inferred from the dataset (v) MISSING_ROGUE included the lowest number of genes (10 of 20), and the lowest percentage of missing data (66%) after removal of rogue taxa (Figs 2, 3). Comparing six datasets, only (v) and (iv) support a monophyletic Tabanomorpha in 50% Majority-rule consensus trees (Fig. 2). Dataset (v) has 664 recovered nodes, and dataset (iv) has 640 recovered nodes while both have same number of taxa. Therefore, we used dataset (v) MISSING_ROGUE as the representative tree for our study.

Aside from reducing overall missing data, we also examined the effects of removing genes that lacked broad taxonomic coverage (inclusion required coverage across at least three infraorders). The COVERAGE dataset (iii) included 14 out of 20 genes, and yielded similarly low bootstrap support for major clades as those that had less missing data [dataset (ii), MISSING]. Yet analyses subsequent to rogue taxa removal showed that minimizing missing data improved bootstrap support substantially more than maximizing taxonomic coverage (Table S1).

Overall, removing rogue taxa from all datasets had the highest impact on improving resolution, support values and phylogenetic decisiveness across the brachyceran tree (Tables 2, 3).

The use of supermatrices for phylogenetic resolution of ancient, hyperdiverse radiations

This study shows both the potential and the limits of traditional Sanger-based molecular systematic datasets that include relatively few genes (~20 loci) to resolve deep phylogeny within a hyperdiverse taxon. With few genes and a high percentage of missing data, the most challenging radiations may resist resolution no matter how dense the taxon sample (see Dell'Ampio *et al.*, 2014). Like previous analyses of such relationships, we find plausible suggestions of phylogenetic history, but many questions remain due to the complexity of character systems, high diversity and ancient rapid radiations. However, the compilation of these independently derived datasets from multiple sources into a comprehensive study provides an opportunity to explore the impacts of taxon sampling, missing data, data distribution and rogue taxa on tree topologies and statistical support. We improved robustness of the inferred phylogenetic relationships by removing rogues and poorly sampled genes from the all-inclusive supermatrix.

Molecular phylogenetic studies have reached a turning point where Sanger sequencing-based methods may often no longer be a cost-effective alternative to next-generation sequencing methods (Trautwein *et al.*, 2012; Yeates *et al.*, 2016). Parallel developments in bioinformatics, computational speed and high-performance computing have changed large-scale phylogenetic analysis. Most recently, analyses based on large-scale genomic data have been successful in resolving controversial problems within insect evolution (e.g. Misof *et al.*, 2014; Peters *et al.*, 2017). Our supermatrix analysis is a benchmark for non-cyclorrhaphan brachyceran phylogeny and a precedent for phylogenomic studies that rely on transcriptome sequencing and genome reduction methods such as anchored hybrid enrichment (e.g. Young *et al.*, 2016) that are now underway.

 doi: 10.1111/syen.12275

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: 10.1111/syen.12275

Figure S1. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset (i), ALL.

Figure S2. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset (ii), MISSING.

Figure S3. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset (iii), COVERAGE.

Figure S4. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset (iv), ALL_ROGUE.

Figure S5. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset (v), MISSING_ROGUE.

Figure S6. Best ML tree with branch support from 100 thorough bootstrap replicates for dataset vi, COVERAGE_ROGUE.

Figure S7. Best ML tree with branch support from 100 rapid bootstrap replicates with outgroup Nematocera.

Table S1. Gene and taxon coverage for dataset (i) with species name, rogue taxa selection list for all six datasets.

Acknowledgements

We thank B. Cassel (North Carolina State University), who assisted with laboratory work and data management for many of the previously published projects, and M. Irwin (University of Illinois at Urbana-Champaign), K. Moulton (The University of Tennessee), S. Morita (Smithsonian National Museum of Natural History), C. Brammer (North Carolina Museum of Natural Sciences), A. Tothova (Masaryk University), K. Collins (Fullerton College), M. Bernasconi (University of Zurich), P. Kerr, S. Gaimari and M. Hauser (California Department of Food and Agriculture) for publishing valuable sequences on NCBI. We also thank K. Meusemann (University of Freiburg) for valuable advice and comments concerning analyses. This project was supported by US National Science Foundation (DEB1257960) and the Doolin Foundation for Biodiversity.

References

Aczél, M.L. (1954) Orthopyga and Campylopyga, new divisions of Diptera. *Annals of the Entomological Society of America*, **47**, 75–80. https://doi.org/10.1093/aesa/47.1.75.

Bocak, L., Barton, C., Crampton-Platt, A., Chesters, D., Ahrens, D. & Vogler, A.P. (2014) Building the Coleoptera tree-of-life for > 8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology*, **39**, 97–110. https://doi.org/10.1111/syen.12037.

Brower, A.V.Z. (2017) Going rogue. *Cladistics*. https://doi.org/10.1111/cla.12211.

Brammer, C.A. & von Dohlen, C.D. (2007) Evolutionary history of Stratiomyidae (Insecta: Diptera): the molecular phylogeny of a diverse family of flies. *Molecular Phylogenetics and Evolution*, **43**, 660–673. https://doi.org/10.1016/j.ympev.2006.09.006.

Brammer, C.A. & von Dohlen, C.D. (2010) Morphological phylogeny of the variable fly family Stratiomyidae (Insecta, Diptera). *Zoologica Scripta*, **39**, 363–377. https://doi.org/0.1111/j.1463-6409.2010.00430.x.

Brauer, F. (1880) I.2. Bemerkungen zur Systematik der Dipteren. Die Zweiflügler des Kaiserlichen Museums zu Wien. *Denkschriften der Kaiserlichen Akademie der Wissenschaften, Wien*, **42**, 105–216.

Brauer, F. (1883) III. Systematische Studien auf Grundlage der Dipteren-Larven nebst einer Zusammenstellung von Beispielen aus der Literatur über dieselben und Beschreibung neuer Formen, Die Zweiflügler des Kaiserlichen Museums zu Wien. *Denkschriften der Kaiserlichen Akademie der Wissenschaften, Wien*, **47**, 1–100.

Burleigh, J.G., Kimball, R.T. & Braun, E.L. (2015) Building the avian tree of life using a large-scale, sparse supermatrix. *Molecular Phylogenetics and Evolution*, **84**, 53–63. https://doi.org/10.1016/j.ympev.2014.12.003.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. & Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287. https://doi.org/10.1126/science.1123061.

Chesters, D. (2016) Construction of a species-level tree-of-life for the insects and utility in taxonomic profiling. *Systematic Biology*, **66**, 426–439. https://doi.org/10.1093/sysbio/syw099.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U. *et al.* (2014) Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution*, **31**, 239–249. https://doi.org/10.1093/molbev/mst196.

Dikow, T. (2009a) A phylogenetic hypothesis for Asilidae based on a total evidence analysis of morphological and DNA sequence data (Insecta: Diptera: Brachycera: Asiloidea). *Organisms Diversity & Evolution*, **9**, 165–188. https://doi.org/10.1016/j.ode.2009.02.004.

Dikow, T. (2009b) Phylogeny of Asilidae inferred from morphological characters of imagines (Insecta: Diptera: Brachycera: Asiloidea). *Bulletin of the American Museum of Natural History*, **319**, 1–175. https://doi.org/10.1206/603.1.

Gouy, M., Guindon, S. & Gascuel, O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221–224. https://doi.org/10.1093/molbev/msp259.

Grimaldi, D.A. (2016) Diverse Orthorrhaphan flies (Insecta: Diptera: Brachycera) in amber from the Cretaceous of Myanmar: *Brachycera in Cretaceous Amber*, *Part VII*. *Bulletin of the American Museum of Natural History*, **408**, 1–31. https://doi.org/10.1206/0003-0090-408.1.1.

Grimaldi, D.A. & Cumming, J.M. (1999) Brachyceran Diptera in Cretaceous ambers and Mesozoic diversification of the Eremoneura. *Bulletin of the American Museum of Natural History*, **239**, 1–124.

Hedtke, S.M., Patiny, S. & Danforth, B.N. (2013) The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. *BMC Evolutionary Biology*, **13**, 138. https://doi.org/10.1186/1471-2148-13-138.

Hennig, W. (1973) 31. Ordnung Diptera (Zweiflügler). *Handbuch der zoologie* (ed. by J.G. Helmcke, D. Starck and H. Wermuth), Vol. 4(2), pp. 1–337. Walter de Gruyter, Berlin.

Hinchliff, C.E. & Roalson, E.H. (2013) Using supermatrices for phylogenetic inquiry: an example using the sedges. *Systematic Biology*, **62**, 205–219. https://doi.org/10.1093/sysbio/sys088.

Irwin, M.E. & Wiegmann, B.M. (2001) A review of the southern African genus *Tongamya* (Diptera: Asiloidea: Mydidae: Megascelinae), with a molecular assessment of the phylogenetic placement of *Tongamya* and the Megascelinae. *African Invertebrates*, **42**, 225–253.

Johnson, S.D. & Morita, S.I. (2006) Lying to Pinocchio: floral deception in an orchid pollinated by long proboscid flies. *Botanical Journal of the Linnean Society*, **152**, 271–278. https://doi.org/10.1111/j.1095-8339.2006.00571.x.

Jønsson, K.A., Fabre, P.H., Kennedy, J.D., Holt, B.G., Borregaard, M.K., Rahbek, C. & Fjeldså, J. (2016) A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides). *Molecular Phylogenetics and Evolution*, **94**, 87–94. https://doi.org/10.1016/j.ympev.2015.08.020.

Karolyi, F., Szucsich, N.U., Colville, J.F. & Krenn, H.W. (2012) Adaptations for nectar feeding in the mouthparts of long proboscid flies (Nemestrinidae: *Prosoeca*). *Biological Journal of the Linnean Society*, **107**, 414–424. https://doi.org/10.1111/j.1095-8312.2012.01945.x.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780. https://doi.org/10.1093/molbev/mst010.

Kergoat, G.J., Soldati, L., Clamens, A.L. *et al.* (2014) Higher level molecular phylogeny of darkling beetles (Coleoptera: Tenebrionidae). *Systematic Entomology*, **39**, 486–499. https://doi.org/10.1111/syen.12065.

Lambkin, C.L., Sinclair, B.J., Pape, T. *et al.* (2013) The phylogenetic relationships among infraorders and superfamilies of Diptera based on morphological evidence. *Systematic Entomology*, **38**, 164–179. https://doi.org/10.1111/j.1365-3113.2012.00652.x.

Lameere, A. (1906) Notes pour la classification des Diptères. *Mémoires de la Société Entomologique de Belgique*, **12**, 105–140.

Lessard, B.D., Cameron, S.L., Bayless, K.M., Wiegmann, B.M. & Yeates, D.K. (2013) The evolution and biogeography of the austral horse fly tribe Scionini (Diptera: Tabanidae: Pangoniinae) inferred from multiple mitochondrial and nuclear genes. *Molecular Phylogenetics and Evolution*, **68**, 516–540. https://doi.org/10.1016/j.ympev.2013.04.030.

van der Linde, K., Houle, D., Spicer, G.S. & Steppan, S.J. (2010) A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genetics Research*, **92**, 25–38. https://doi.org/10.1017/S001667231000008X.

McMahon, M.M. & Sanderson, M.J. (2006) Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Systematic Biology*, **55**, 818–836. https://doi.org/10.1080/10635150600999150.

Melander, A.L. (1950) Taxonomic notes on some smaller Bombyliidae (Diptera). *Pan-Pacific Entomologist*, **26**, 139–156.

Meredith, R.W., Janecka, J.E., Gatesy, J. *et al.* (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science*, **334**, 521–524. https://doi.org/10.1126/science.1211028.

Misof, B., Meyer, B., von Reumont, B., Kück, P., Misof, K. & Meusemann, K. (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics*, **14**, 348. https://doi.org/10.1186/1471-2105-14-348.

Misof, B., Liu, S., Meusemann, K. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767. https://doi.org/10.1126/science.1257570.

Morita, S.I., Bayless, K.M., Yeates, D.K. & Wiegmann, B.M. (2016) Molecular phylogeny of the horse flies: a framework for renewing tabanid taxonomy. *Systematic Entomology*, **41**, 56–72. https://doi.org/10.1111/syen.12145.

Pape, T., Blagoderov, V. & Mostovski, M.B. (2011) Order Diptera Linnaeus, 1758. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (ed. by Z.Q. Zhang). *Zootaxa*, **3148**, 222–229.

Peters, R.S., Meyer, B., Krogmann, L. *et al.* (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biology*, **9**, 55. https://doi.org/10.1186/1741-7007-9-55.

Peters, R.S., Krogmann, L., Mayer, C. *et al.* (2017) Evolutionary history of the Hymenoptera. *Current Biology*, **27**, 1013–1018. https://doi.org/10.1016/j.cub.2017.01.027.

Pirie, M.D., Humphreys, A.M., Galley, C. *et al.* (2008) A novel supermatrix approach improves resolution of phylogenetic relationships in a comprehensive sample of danthonioid grasses. *Molecular Phylogenetics and Evolution*, **48**, 1106–1119. https://doi.org/10.1016/j.ympev.2008.05.030.

Piwczyński, M., Szpila, K., Grzywacz, A. & Pape, T. (2014) A large-scale molecular phylogeny of flesh flies (Diptera: Sarcophagidae). *Systematic Entomology*, **39**, 783–799. https://doi.org/10.1111/syen.12086.

de Queiroz, A. & Gatesy, J. (2007) The supermatrix approach to systematics. *Trends in Ecology & Evolution*, **22**, 34–41. https://doi.org/10.1016/j.tree.2006.10.002.

Sanderson, M.J., McMahon, M.M. & Steel, M. (2010) Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evolutionary Biology*, **10**, 155. https://doi.org/10.1186/1471-2148-10-155.

Sinclair, B.J. & Cumming, J.M. (2006) The morphology, higher-level phylogeny and classification of the Empidoidea (Diptera). *Zootaxa*, **1180**, 1–172.

Sinclair, B.J., Cumming, J.M. & Brooks, S.E. (2013) Male terminalia of Diptera (Insecta): a review of evolutionary trends, homology and phylogenetic implications. *Insect Systematics & Evolution*, **44**, 373–415. https://doi.org/10.1163/1876312X-04401001.

Smith, S.A., Beaulieu, J.M. & Donoghue, M.J. (2009) Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology*, **9**, 37. https://doi.org/10.1186/1471-2148-9-37.

Soltis, D.E., Smith, S.A., Cellinese, N. *et al.* (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany*, **98**, 704–730. https://doi.org/10.3732/ajb.1000404.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Steel, M. & Sanderson, M.J. (2010) Characterizing phylogenetically decisive taxon coverage. *Applied Mathematics Letters*, **23**, 82–86. https://doi.org/10.1016/j.aml.2009.08.009.

Stuckenberg, B.R. (2001) Pruning the tree: a critical review of the classification of the Homeodactyla (Diptera, Brachycera), with new perspectives and an alternative classification. *Studia Dipterologica*, **8**, 3–41.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729. https://doi.org/10.1093/molbev/mst197.

Thomson, R.C. & Shaffer, H.B. (2010) Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Systematic Biology*, **59**, 42–58. https://doi.org/10.1093/sysbio/syp075.

Trautwein, M.D., Wiegmann, B.M. & Yeates, D.K. (2010) A multigene phylogeny of the fly superfamily Asiloidea (Insecta): taxon sampling and additional genes reveal the sister-group to all higher flies (Cyclorrhapha). *Molecular Phylogenetics and Evolution*, **56**, 918–930. https://doi.org/10.1016/j.ympev.2010.04.017.

Trautwein, M.D., Wiegmann, B.M. & Yeates, D.K. (2011) Overcoming the effects of rogue taxa: evolutionary relationships of the bee flies. *PLOS Currents*, **3**, RRN1233. https://doi.org/10.1371/currents.RRN1233.

Trautwein, M.D., Wiegmann, B.M., Beutel, R., Kjer, K.M. & Yeates, D.K. (2012) Advances in insect phylogeny at the dawn of the postgenomic era. *Annual Review of Entomology*, **57**, 449–468. https://doi.org/10.1146/annurev-ento-120710-100538.

Vaidya, G., Lohman, D.J. & Meier, R. (2011) SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, **27**, 171–180. https://doi.org/10.1111/j.1096-0031.2010.00329.x.

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. https://doi.org/10.1080/10635150390250965.

Wiegmann, B.M., Trautwein, M.D., Winkler, I.S. *et al.* (2011) Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 5690–5695. https://doi.org/10.1073/pnas.1012675108.

Winterton, S.L. & Ware, J.L. (2015) Phylogeny, divergence times and biogeography of window flies (Scenopinidae) and the therevoid clade (Diptera: Asiloidea). *Systematic Entomology*, **40**, 491–519. https://doi.org/10.1111/syen.12117.

Winterton, S.L., Yang, L.L., Wiegmann, B.M. & Yeates, D.K. (2001) Phylogenetic revision of Agapophytinae subf.n. (Diptera : Therevidae) based on molecular and morphological evidence. *Systematic Entomology*, **26**, 173–211. https://doi.org/10.1046/J.1365-3113.2001.00142.X.

Winterton, S.L., Wiegmann, B.M. & Schlinger, E.I. (2007) Phylogeny and Bayesian divergence time estimations of small-headed flies (Diptera: Acroceridae) using multiple molecular markers. *Molecular Phylogenetics and Evolution*, **43**, 808–832. https://doi.org/10.1016/j.ympev.2006.08.015.

Woodley, N.E. (1989) Phylogeny and classification of the "Orthorrhaphous" Brachycera. *Manual of Nearctic Diptera* (ed. by J.F. McAlpine), pp. 1371–1395. Canadian Government Publishing Centre, Hull, Ottawa.

Woodley, N.E., Borkent, A. & Wheeler, T.A. (2009) 5 Phylogeny. *Manual of Central American Diptera* (ed. by B.V. Brown, A. Borkent, J.M. Cumming, D.M. Wood, N.E. Woodley and M. Zumbado). NRC Research Press.

Yang, L.L., Wiegmann, B.M., Yeates, D.K. & Irwin, M.E. (2000) Higher-level phylogeny of the Therevidae (Diptera : Insecta) based on 28S ribosomal and elongation factor-1 alpha gene sequences. *Molecular Phylogenetics and Evolution*, **15**, 440–451. https://doi.org/10.1006/Mpev.1999.0771.

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

Yeates, D.K. (2002) Relationships of extant early Brachycera (Diptera): a quantitative synthesis of morphological characters. *Zoologica Scripta*, **31**, 105–121. https://doi.org/10.1046/j.0300-3256.2001.00077.x.

Yeates, D.K. & Greathead, D. (1997) The evolutionary pattern of host use in the Bombyliidae (Diptera): a diverse family of parasitoid flies. *Biological Journal of the Linnean Society*, **60**, 149–185. https://doi.org/10.1111/j.1095-8312.1997.tb01490.x.

Yeates, D.K. & Irwin, M.E. (1996) Apioceridae (Insecta: Diptera): cladistic reappraisal and biogeography. *Zoological Journal of the Linnean Society*, **116**, 247–301.

Yeates, D.K. & Wiegmann, B.M. (1999) Congruence and controversy: toward a higher-level phylogeny of Diptera. *Annual Review of Entomology*, **44**, 397–428. https://doi.org/10.1146/annurev.ento.44.1.397.

Yeates, D.K., Meusemann, K., Trautwein, M., Wiegmann, B.M. & Zwick, A. (2016) Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution. *Current Opinion in Insect Science*, **13**, 16–23. https://doi.org/10.1016/j.cois.2015.10.007.

Young, A.D., Lemmon, A.R., Skevington, J.H. *et al.* (2016) Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evolutionary Biology*, **16**, 143. https://doi.org/10.1186/s12862-016-0714-0.

Accepted 26 September 2017