A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families


1 Department of Biological Sciences, National University of Singapore, Singapore, Singapore, 2 Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, MO, U.S.A., 3 Biology I, Evolutionary Biology & Ecology, University of Freiburg, Freiburg (Brsg.), Germany, 4 Center for Molecular Biodiversity Research, Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany, 5 Australian National Insect Collection, CSIRO National Research Collections Australia (NRCA), Canberra, Australia and 6 Evolution and Ecology, Research School of Biological Sciences, The Australian National University, Canberra, Australia

Abstract. Culicomorpha is a particularly species-rich clade within Diptera (true flies) that comprises c. 10% of the described diversity, including many medically important flies. Morphological studies – even when all life stages are included – yield relationships different from those derived from molecular data, notably with regard to the position of Chironomidae. Congruence amongst molecular studies has been weak due to limitations in gene- and family-level taxon coverage. Here we use a whole-transcriptome shotgun phylogenomic approach to clarify the relationships among all families of Culicomorpha. The dataset comprised 30 species (27 ingroup) and 364 888 amino acid residues for 1233 single-copy protein-encoding genes. Likelihood and parsimony analyses produce robust and highly congruent phylogenetic trees, with only one node in conflict. The superfamily Culicoidea is well supported and comprises Dixidae + [Corethrellidae + (Chaoboridae + Culicidae)]. As suggested previously, Chironomoidea is not monophyletic. The well supported Thaumaleidae + Simuliidae is sister group to Culicoidea, with the weakly supported Chironomidae + Ceratopogonidae probably being the sister group of all remaining Culicomorpha. We used random addition concatenation analysis (RADICAL) and four-cluster likelihood mapping (FcLM) to assess the strengths of nodal support. The sister-group relationship between Chironomidae + Ceratopogonidae is consistent with the FcLM results but support for this relationship emerges only when 1150 of the 1233 loci are analysed. We discuss briefly nodes that remain poorly supported even with thousands of genes and mention problems with vouchering in transcriptomic studies.

Introduction

Culicomorpha is a monophyletic group within Diptera (true flies) and includes a diversity of medically important biting flies belonging to eight family-level taxa. The monophyly of all families, namely the Ceratopogonidae (biting midges, no-see-ums; Fig. 1A–C), Chironomidae (non-biting midges; Fig. 1D–G), Thaumaleidae (trickle or solitary midges; Fig. 1H, I), Simuliidae (black flies, punkies, buffalo gnats or ‘sand flies’; Fig. 1J–L), Dixidae (meniscus midges; Fig. 1M, N), Corethrellidae (frog-biting midges; Fig. 1O), Chaoboridae (phantom midges; Fig. 1P, Q) and Culicidae (mosquitoes; Fig. 1R–T), is well supported by morphological characters derived from all life stages, as summarized by Borkent (2012). Some families are species-rich (Culicidae, Chironomidae,
Fig. 1. Examples of living Culicomorpha – tribe/subfamily in parentheses where appropriate. All photographs © Steve Marshall. (A) Ceratopogonidae; adult female, *Culicoides* sp., (Ceratopogoninae); (B) Ceratopogonidae; adult female, *Dasyhelea* sp. (Dasyheleinae); (C) Ceratopogonidae; larva, *Polypomia* sp.; (D) Chironomidae; adult male and female, *Cricotopus trifasciatus* Meigen (Orthocladiinae); (E) Chironomidae; adult male, *Diamesa* sp. (Diamesinae); (F) Chironomidae; pupa, *Ablabesmyia* sp. (Tanytopinae); (G) Chironomidae; larva, *Chironomus* sp. (Chironominae); (H) Thaumaeidae; adult, *Androprosopograpta* sp.; (I) Thaumaeidae; larva, *Thaumalea* sp.; (J) Simuliidae; adult female, *Simulium bicoloratum* Malloch; (K) Simuliidae; pupal cast skin, *Simulium venustum* Say; (L) Simuliidae; larval head, *Simulium vittatum* Zetterstedt; (M) Dixidae; adult, *Dixella cornuta* Johannsen; (N) Dixidae; larva, *Notodixa* sp.; (O) Corethrellidae; adult female, *Corethrella wirthei* Stone; (P) Chaoboridae; adult female, *Chaoborus* sp.; (Q) Chaoboridae; larva, *Chaoborus* sp.; (R) Culicidae; adult female, *Anopheles* sp. (Anophelineae); (S) Culicidae; adult male, *Toxorhynchites* sp., (Toxorhynchitini); (T) Culicidae; pupa, larva, *Culex territans* Walker (Culicini). [Colour figure can be viewed at wileyonlinelibrary.com].
Ceratopogonidae each with several thousand described species) and include important vectors because adult females of four groups include species that feed on vertebrate blood (Culicidae, Simuliidae, Ceratopogonidae and Corethrellidae). Various species in these groups are capable of transmitting pathogens, some of which cause diseases of humans and livestock, including dengue, malaria, yellow fever, onchocerciasis and bluetongue (Gullan & Cranston, 2014). Almost without exception culicomorphans have aquatic immature stages and these provide valuable evidence in systematics and in ecological studies. For example, the diversity and dominance of larvae of Chironomidae in freshwater systems encourage their widespread use for bioassessment of aquatic ecosystem health, and monitoring of larval Culicidae and Simuliidae is vital to surveillance of disease vectors (Gullan & Cranston, 2014).

Culicomorpha is one of the higher-level groupings in Diptera that have long been recognized as a natural taxon. For example, even prior to explicit matrix-based phylogenetics, Crompton (1924, 1925) proposed the group and discussed relationships based on adult morphology, including details of the complex thoracic sclerites. Edwards (1926) overall agreed with Crompton’s views adding that all life stages (i.e. egg, larva, pupa and adult of both sexes) should be used in deducing phylogenies. These authors pioneered character-based arguments antecedent to Hennig’s (1950, 1966) formalized phylogenetic systematics. Clear statements were made concerning differentiation of ancestral from derived features, and of morphological convergence. However, it was left to Hennig (1973) to propose Culicomorpha formally as a taxon composed of the aforementioned eight families. Regarding internal relationships, Hennig (1973) proposed two superfamilies (Fig. 2A). The Culicoidea comprised Dixidea as sole representative of family group Dixidea, with the Culicidae and Chaoboridae (including Corethrella Coquillett, now a separate family; Wood & Borkent, 1989), forming the family group Culicidea. According to Hennig, the sister group of Culicoidea is the superfamily Chironomoidea, comprising Thaumaleidae, Simuliidae, Ceratopogonidae and Chironomidae, with Thaumaleidae allocated to family group Thaumaleidea and the remaining three families to Chironomoidea.

Since Hennig’s proposal, a monophyletic Culicomorpha has been supported by morphological and molecular data, but its sister group remains elusive. Lambkin et al. (2013), in a morphological analysis, suggested Culicomorpha as sister to all other Diptera except for the Nymphomyiidae, which was proposed to be sister group to all Diptera (including Culicomorpha). The enigmatic Nymphomyiidae (with or without Axymyiidae according to analysis) has been suggested as sister to Culicomorpha (e.g. Bertone et al., 2008) and has even been discussed occasionally as being included within the Culicomorpha (Hennig, 1981; Sæther, 2000). Identification of the sister group of Culicomorpha using molecular data has been hindered by sparse gene and taxon sampling (Pawlowski et al., 1996; Bertone et al., 2008). Inadequate gene and taxon sampling may also explain unexpected results such as the suggestion that the Axymyiidae alone (Sevčík et al., 2016) or in combination with Nymphomyiidae (Bertone et al., 2008) are the sister group of Culicomorpha. Based on these results, it appears likely that resolving the sister group to Culicomorpha will require large-scale analyses of all family-level taxa across the nematoceran dipteran grade.

Our understanding of relationships within Culicomorpha is fairly typical for infraorder-level taxa in that some clades are firmly established while other family-level taxa are difficult to place (e.g. Calyptratae: Kutty et al., 2010). For example, Culicoida finds consistent support in traditional ‘hennigian’ (Borkent, 2012) (Fig. 2H), parsimony-based studies of morphology (e.g. Wood & Borkent, 1989; Oosterbroek & Court- ney, 1995) (Fig. 2B), molecular studies (Bertone et al., 2008; Fig. 2F), and combined analyses (Fig. 2G). The Culicoida include Corethrellidae, Chaoboridae and Culicidae, with the latter being sister groups and Dixidae sister group to these three. However, these relationships are not recovered in some studies with sparse gene and taxon sampling (28S ribosomal RNA; Pawlowski et al., 1996; Fig. 2C; small subunit ribosomal RNA, Miller et al., 1997 (Fig. 2I), Aransay et al., 2000; ribosomal protein C-terminal extension, Fallon & Li, 2007). Uniquely amongst published studies, Wiegmann et al., 2011 proposed the sister group to Culicidae as Corethrellidae + Chaoboridae (Fig. 2G) in an analysis based on 14 nuclear genes, full mitochondrial genomes, and 371 morphological characters, although the matrix was incomplete for several culicomorphans.

In contrast to the relative stability of Culicoidea, the composition and internal arrangement of the families of the Chironomoidea have been volatile (see summary in Sæther, 2000 and Fig. 2). The monophyly of the Chironomoidea is recovered only occasionally and the constituent families are more likely to form a paraphyletic grade. Major uncertainty concerns the placement of Thaumaleidae, which has been hypothesized to be the sister group to Culicomorpha or even excluded from Culicomorpha (with or without Nymphomyiidae) by Krzeminska et al. (1993) in a noncladistic analysis. When recovered within Chironomoidea, it has been inferred as the sister family to Simuliidae in molecular studies (Moulton, 2000; Wiegmann et al., 2011; Cranston et al., 2012) and in a morphological synthesis by Borkent (2012). This relationship conflicts with previous placements based on morphological evidence by Hennig (1973) (Fig. 2A), Wood & Borkent (1989) (Fig. 2B), Oosterbroek & Courtney (1995) (Fig. 2B) and Sæther (2000) (Fig. 2D, E). The estimated relationships between the remaining family-level taxa in Chironomoidea vary (see Fig. 2). Not even the sister-group relationship of the two main midge families, the Chironomidae and Ceratopogonidae, is uncontroversial. This relationship, proposed by Edwards (1926) and Hennig (1973) (Fig. 2A), was supported subsequently by Wood & Borkent (1989) and Oosterbroek & Courtney (1995) (Fig. 2B) based on morphology. It is recovered in the molecular analyses of Bertone et al. (2008)(Fig. 2F) and Wiegmann et al. (2011)(Fig. 2G), but not supported by Sæther’s (2000) morphological studies (Fig. 2D, E). Borkent (2012) expressly discusses this relationship and, although he recovered Simulioidae, finds the Ceratopogonidae to be sister group of Thaumaleidae + Simuliidae, and not of Chironomidae (Fig. 2H).

Incongruence between the studies outlined may be due to both undersampling of taxa and genes and some inadequate analytical
Fig. 2. Previous hypotheses of relationships within infraorder Culicomorpha. (A) Hennig (1973) (Corethrella treated as a Chaoboridae)(morphology); (B) Wood & Borkent (1989), Oosterbroek & Courtney (1995) (morphology); (C) Pawlowsky et al. (1996) (molecular data); (D, E) Sæther (2000) two ‘common’ trees (morphology); (F) Bertone et al. (2008) (molecular data); (G) Wiegmann et al. (2011) (molecular data); (H) Borkent (2012) (morphology); (I) Miller et al. (1997) (molecular data). The counts and branch thicknesses indicate the number of times the relationship is present in other topologies illustrated here. Dashed lines (- - -) denote a unique relationship hypothesis across topologies illustrated in this figure. [Colour figure can be viewed at wileyonlinelibrary.com].

approaches, compounded by uncertainty concerning the close relatives of the Culicomorpha which would help with determining homologies and character polarities. Here we address the sampling problems of molecular studies by including species from all family-level taxa and using a phylogenomic approach based on high-throughput sequencing of transcriptomes for 19 Culicomorpha species. After combining the new data with published transcriptomes, our dataset includes 27 species, i.e. our study greatly enhances the available molecular sequence data for Culicomorpha, and a wide range of analyses yields largely consistent and well-supported relationship hypotheses.

Materials and methods

Ethics statement

Specimens were collected widely, under collection permits issued by appropriate authorities where required. No species are endangered or protected in any part of their distribution.

Taxa. Material was collected by conventional field methods for the (largely) aquatic taxa. We included a large range of subfamilies (Table 1) whenever possible. Where feasible, specimens of immature stages were viewed live with a stereomicroscope in the field to assess identity (e.g. for subfamilies of Chironomidae), then immersed alive in RNAlater (Qiagen, Valencia, CA, U.S.A.) and crushed with a sterile pestle. Samples were kept cold until storage at −80°C. Several species were obtained as immature stages from pitchers of Nepenthes ampullaria (Jack (1835)) growing in Singapore. Some larvae were reared to adulthood by feeding them with Caenorhabditis elegans (Maupas, 1900) (Corethrellidae, Ceratopogonidae and small Culicidae) or insect larvae, organic matter or microbes from the same pitcher (see Table 1: Toxorhynchites sp., Tripteroides aranoides, Corethrella calathicola and Dasyhelea sp.). Whenever feasible, body parts essential to identification were retained as slide mounts prepared according to standard methods. In addition, some specimens were identified to species level via DNA barcoding using cytochrome c oxidase subunit 1.
### Table 1. List of newly added taxa, codes, life stages, locations.

<table>
<thead>
<tr>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
<th>No.</th>
<th>Stage/sex</th>
<th>Locality</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Date</th>
<th>Collector</th>
<th>Determined</th>
<th>Identification reference or source</th>
<th>Author names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scatopsidae</td>
<td>Indet</td>
<td>indet</td>
<td>2</td>
<td>m</td>
<td>Goodradigbe, Australia</td>
<td>35.22°S</td>
<td>148.49°E</td>
<td>29.vi.2016</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td></td>
</tr>
<tr>
<td>Ceratopogonidae</td>
<td>Dasyhelea</td>
<td>sp.</td>
<td>1</td>
<td>ad from L</td>
<td>Neesoon, Singapore</td>
<td>1.23°N</td>
<td>103.48°E</td>
<td>ii.2015</td>
<td>Wing</td>
<td>Wing/Borkent</td>
<td>Comparison with collection material</td>
<td>Kieffer, 1911</td>
</tr>
<tr>
<td>Ceratopogonidae</td>
<td>Atrichopogon</td>
<td>sp.</td>
<td>1</td>
<td>ad/m, f</td>
<td>Goodradigbe, Australia</td>
<td>35.22°S</td>
<td>148.49°E</td>
<td>29.vi.2016</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>Kieffer, 1906</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Pedronomas</td>
<td>sp.</td>
<td>2</td>
<td>P</td>
<td>Canberra, Australia</td>
<td>33.31°S</td>
<td>148.56°E</td>
<td>28.vi.2016</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>Philippi, 1865</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Trissopelia</td>
<td>nemorum</td>
<td>2</td>
<td>L</td>
<td>Cornwall, England</td>
<td>50.34°N</td>
<td>4.41°W</td>
<td>19.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Goethebuer, 1921)</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Chironomidae</td>
<td>Podonomus</td>
<td>2</td>
<td>P</td>
<td>Canberra, Australia</td>
<td>35.31°S</td>
<td>148.56°E</td>
<td>9.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Skuse, 1889)</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Chironomidae</td>
<td>Atrichopogon</td>
<td>4</td>
<td>P</td>
<td>Pandan Res, Singapore</td>
<td>50.13°N</td>
<td>103.74°E</td>
<td>17.vii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Dobyns, 1912)</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Chironomidae</td>
<td>Polypedilum</td>
<td>1</td>
<td>P</td>
<td>Canberra, Australia</td>
<td>35.31°S</td>
<td>148.56°E</td>
<td>9.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Kieffer, 1912)</td>
</tr>
<tr>
<td>Chironomidae</td>
<td>Chironomidae</td>
<td>Kiefferophyes</td>
<td>3</td>
<td>P</td>
<td>Canberra, Australia</td>
<td>33.31°S</td>
<td>148.56°E</td>
<td>28.vi.2016</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Skuse, 1889)</td>
</tr>
<tr>
<td>Simuliidae</td>
<td>Paraheptagyia</td>
<td>tonnoiri</td>
<td>2</td>
<td>P</td>
<td>Canberra, Australia</td>
<td>35.31°S</td>
<td>148.56°E</td>
<td>9.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>Rubesov, 1962</td>
</tr>
<tr>
<td>Simuliidae</td>
<td>Simulium</td>
<td>sp.</td>
<td>1</td>
<td>P</td>
<td>Cornwall, England</td>
<td>50.34°N</td>
<td>4.41°W</td>
<td>19.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>(Theobald, 1902)</td>
</tr>
<tr>
<td>Dixidae</td>
<td>Dixia</td>
<td>sp.</td>
<td>2</td>
<td>L</td>
<td>Cornwall, England</td>
<td>50.34°N</td>
<td>4.41°W</td>
<td>19.iii.2015</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>Meigen, 1818</td>
</tr>
<tr>
<td>Dixidae</td>
<td>Dixella/Nothodixa</td>
<td>sp.</td>
<td>2</td>
<td>L</td>
<td>Canberra, Australia</td>
<td>35.22°S</td>
<td>148.49°E</td>
<td>29.vi.2016</td>
<td>Cranston</td>
<td>Cranston</td>
<td>Comparison with collection material</td>
<td>Dyer &amp; Shannon, 1924</td>
</tr>
<tr>
<td>Corethrellida</td>
<td>Corethrella</td>
<td>calathicola</td>
<td>1</td>
<td>ad from L</td>
<td>Neesoon, Singapore</td>
<td>1.23°N</td>
<td>103.48°E</td>
<td>ii.2015</td>
<td>Wing</td>
<td>Borkent</td>
<td>Comparison with collection material</td>
<td>Edwards, 1930</td>
</tr>
<tr>
<td>Chaoboridae</td>
<td>Chaoborus</td>
<td>floridulus</td>
<td>1</td>
<td>ad</td>
<td>Upper Seletar Res, Singapore</td>
<td>1.40°N</td>
<td>103.80°E</td>
<td>ii.2015</td>
<td>Wing</td>
<td>Wing</td>
<td>Comparison with collection material</td>
<td>Edwards, 1930</td>
</tr>
<tr>
<td>Culicidae</td>
<td>Toxorhynchites</td>
<td>sp.</td>
<td>1</td>
<td>ad from L</td>
<td>Neesoon, Singapore</td>
<td>1.23°N</td>
<td>103.48°E</td>
<td>ii.2015</td>
<td>Wing/Dr. Ian Mendenhall</td>
<td>COI match to GenBank accession number KP461761 (90%)</td>
<td>Theobald, 1901</td>
<td></td>
</tr>
<tr>
<td>Culicidae</td>
<td>Tripteryges</td>
<td>anansoides</td>
<td>3</td>
<td>ad from L</td>
<td>Neesoon, Singapore</td>
<td>1.23°N</td>
<td>103.48°E</td>
<td>ii.2015</td>
<td>Wing</td>
<td>Wing</td>
<td>COI match to GenBank accession number AF425850 (95%)</td>
<td>(Theobald, 1901)</td>
</tr>
</tbody>
</table>
Specimen vouchers are deposited at the Lee Kong Chian Natural History Museum, Singapore. Total RNA extractions were carried out using Trizol reagent (Invitrogen, ThermoFisher Scientific, MA, U.S.A.) following the manufacturer’s protocol and an additional DNase treatment was included using the Ambion™ TURBO DNA-free™ kit (Invitrogen) following the manufacturer’s protocol and RNA was then pooled from three different adult specimens in the same cDNA library due to insufficient RNA quantity; all other libraries were prepared with RNA of a single specimen from each species. The libraries (insert size 500 bp) were prepared using the Illumina TruSeq Stranded mRNA Library Prep Kit by Abiotech Pte Ltd (Singapore). Paired-end sequencing (250 bp) of eight multiplexed libraries was carried out in one lane of the Illumina HiSeq 2500 Sequencing system at the Genome Institute of Singapore (GIS, Singapore) and the Singapore Centre for Environmental Life Sciences Engineering (SCELSE, Singapore) sequencing facilities. Additionally, sequence data for ten species were downloaded from databases (GenBank SRA, TSA or VectorBase; see Table S1). Raw sequence data generated in this study are deposited in the NCBI SRA database (Table 2).

De novo assemblies. De novo assemblies of the paired-end transcriptome data were generated in CLC GENOMICS WORKBENCH 7.5.1 (https://www.qiagenbioinformatics.com/) for all newly sequenced species as well as two species for which raw data had been downloaded from NCBI. Reads were trimmed (limit 0.001), and the de novo assemblies were carried out under optimized parameters (word size = 50, bubble size = 100, identity fraction = 1, length fraction = 1). Contigs with average coverage of < 10 and a length < 100 bp were discarded, and we applied a final trimming of 10 bp from both ends on each contig sequence. A cross-contamination check across libraries was done for de novo assembled transcripts sequenced in the same lane, an all-versus-all search using blastn (BLAST 2.2.28+; Camacho et al., 2009); contigs that shared high sequence similarity (blastn parameters: length > 200 bp, identity ≥ 98%, mismatches ≤ 4) were excluded from further analyses.

Orthology prediction and design of the dataset. Orthology prediction, post-processing and dataset preparation followed Misof et al. (2014). Orthology assignment was done using ORTHOGRAPH v 0.5.8 (Petersen et al., 2017) based on an orthologue reference set of 3288 orthologous sequences of single-copy protein-encoding genes from five reference species (hierarchy split set to between Hymenoptera and remaining Holometabola): Anopheles gambiae, Tribolium castaneum, Drosophila melanogaster, Mayetiola destructor, and Bombyx mori (Table S2). These single-copy genes were extracted and assembled from OrthoDB5 (Waterhouse et al., 2011). For each single-copy gene, sequence alignment of included reference species and building of profile hidden Markov models (pHMMs) were automatically conducted in ORTHOGRAPH. Contigs from individual species transcriptome assemblies were translated into all six reading frames and searched against the pHMMs of each single-copy gene included in the orthologue reference set. ORTHOGRAPH confirms successful transcript matches (candidate orthologous sequences) with a reciprocal blast search against the protein sequences of the full proteome of any of the five reference species and keeps transcripts for which the best reciprocal hit criterion is fulfilled (minimal transcript length = 30aa, Selenocysteine ‘U’ was always replaced by ‘X’ to avoid problems in downstream analyses; other settings default).
Transcripts successfully assigned to single-copy genes from the orthologue reference on a translational level and corresponding nucleotide gene sets were summarized using custom perl scripts (Peters et al., 2017). Internal stop codons, if any, were replaced with an ‘X’ and with ‘NNN’ on the nucleotide level. The amino acid transcript sequences were aligned using MAFFT v.7.123b (Katoh & Standley, 2013) with the L-INS-i algorithm. The resulting multiple sequence alignments (MSAs) were scrutinized for outliers, and outlier sequences were refined as described in Peters et al. (2017). After a second outlier check, remaining outliers were removed. Subsequently, we removed all sequences of the reference species except for those from A. gambiae and M. destructor (used as outgroup in this study).

Then we generated corresponding nucleotide sequence alignments for all single-copy genes with Pal2Nal (Suyama et al., 2006) using the refined amino acid MSAs as blueprint. Sequence similarity in the amino acid alignments that may be random or ambiguously aligned were identified using ALICORE v. 2.2 (Misof & Misof, 2009; Kück et al., 2010) with the maximal number of pairwise comparisons, using the option -e for gappy alignments derived from RNASeq data and otherwise defaults. All ambiguously aligned identified sections and the corresponding regions in the nucleotide MSAs were removed using custom perl scripts. Before final concatenation of all MSAs into a supermatrix with FASconCAT-G (Kück & Longo, 2014), the leading and trailing gaps in the sequence alignment were recoded as ‘X’ for the amino acids and ‘N’ for nucleotides. All of these steps were carried out using custom perl scripts published and available from http://software.zfmk.de/The-1KITE-project_evolution-of-insects.zip and used in Misof et al. (2014).

To improve the overall information content (IC) of the supermatrix, MARE v0.1.2-rc (Misof et al., 2013) was utilized with taxa weighting ($t = 3$) to retain all the species in the reduced dataset. A corresponding optimized nucleotide dataset was generated accordingly with custom made perl-scripts because MARE is applicable for amino acid datasets only. An ALISTAT (https://github.com/thomaskf/AliStat) analysis (version 1.7) was also carried out on the amino acid datasets to determine the missing data per site as used in Misof et al. (2014).


The partitioned amino acid dataset was analysed in RAxML-HPC v 8.2.9 (Stamatakis, 2014) XSEDE on the CIPRES computer cluster (Cyberinfrastructure for Phylogenetic Research; San Diego Supercomputer Center) (Miller et al., 2010). The search for the best-scoring maximum likelihood (ML) tree (20 multiple inferences) was carried out under the best partition scheme and bestmodel+GAMMA. Multiple bootstrap replicates were carried out with a MRE-based Bootstopping criterion. A random addition concatenation analysis (RADICAL) (Narechania et al., 2012) was also performed on the dataset. Ten randomizations of 50 step loci additions were carried out using both RAxML v.8.2.8 (Stamatakis, 2014) and RAUP (Swofford, 2000) for tree reconstructions, and the fixation point of all node and area under the curve (auc) values were calculated. A four-cluster likelihood mapping (FiLM) analysis (Strimmer & von Haseeler, 1997) implemented in RAxML v.8.2.8 (Stamatakis, 2014) was conducted on the amino acid dataset to test alternative hypotheses for the position of Ceratopogonidae (sister group to Chironomidae or remaining Culicomorpha) (see Table S3 for defined groups and topologies). A custom-made perl script was used to map the results graphically. The results of the ML tree inference were compared with those of parsimony trees, which were obtained with TNT v1.5 (Goloboff & Catalano, 2016) (new technology search, level 10, hits 20, gaps coded as missing data). Node support was assessed by jackknife (JK) resampling (1000 replicates) at 36% deletion.

The nucleotide dataset was analysed using a partitioned likelihood analysis. We also analysed two other reduced nucleotide datasets consisting of: (i) first and second codon positions; (ii) only second codon positions. For the full nucleotide alignment, best-fit partitioning schemes for the genes and the codon positions were established in PARTITIONFINDER2 v 2.1.1 (Stamatakis, 2014; Lanfear et al., 2014; Lanfear et al., 2016) using the GTR and GTR + GAMMA models. Partitioning schemes for the two reduced datasets were determined under the same model. ML searches and bootstrapping with an MRE-based Bootstopping criterion for all three nucleotide datasets were performed in RAxML-HPC v 8.2.9 (Stamatakis, 2014) XSEDE on the CIPRES computer cluster (Miller et al., 2010).

**Results**

We generated shotgun transcriptomic data for 19 species across the eight families allocated to Culicomorpha. Newly collected material specific to this project came from Australia, England and Singapore (Table 1). Sequences for another eight culicomorphan species were obtained from public databases, namely NIAD’s Vectorbase, NCBI GenBank’s Sequence Read Archive (SRA) and Transcriptome Shotgun Assembly (TSA) (Table S1). We used three ‘outgroup’ taxa one from each of three infraorders of the ‘lower’ Diptera: the scatopsid is a newly generated transcriptome, and the cecid and psychodid transcriptomes are from public databases (SI Tables 1, 2). Transcriptomes
Fig. 3. Amino acid maximum likelihood topology (ML) with bootstrap support values and parsimony (MP) jackknife values mapped above nodes. RADICAL results for both likelihood and parsimony-based analyses are shown below nodes. auc, area under the curve. [Colour figure can be viewed at wileyonlinelibrary.com].

for Axomyiidae or Nymphomyiidae, which have been discussed as possible sister taxa to Culicomorpha, were not available.

Table 2 provides information on the included species (19 de novo generated transcriptomes plus 11 from published databases), number of assembled contigs, and number of mapped orthologs. After post-processing (i.e. alignment, alignment refinement, outlier check and outlier removal, identification and deleting ambiguously aligned sections from the MSAs and corresponding processes on nucleotide level), the concatenated supermatrix yielded 3131 single-copy protein-encoding genes. The coverage of the data matrix in terms of gene occurrence was 0.71 (IC = 0.36) and the coverage of the matrix according to each site was 0.63. The final dataset (after optimization with MARE) comprised 30 species and an alignment of 364,888 amino acid sites representing 1233 single-copy protein-encoding genes (coverage of the data matrix in terms of occurrence of genes was 0.85, IC = 0.53, coverage according to sites was 0.74; Fig. S6). The phylogenetic trees from both likelihood and parsimony amino acid analyses were robust and highly congruent, with only one conflicting node (Fig. 3). Nearly all nodes (25 of 27) showed maximal bootstrap support (BS) after achieving bootstrap convergence and parsimony jackknife support (JK) of 100, only two clades showed a low support (ML: Ceratopogonidae + Chironomidae: 0.72 BS; MP: Dixidae placement: 0.92 JK).

The RADICAL analysis on the amino acid dataset revealed that, although BS and JK values were high, certain nodes in the Culicomorpha were difficult to resolve. Nineteen of the 27 nodes could be resolved with as few as 50 loci, whereas one node required as many as 1150 (out of 1233) to place taxa confidently on the tree. Nodes that were particularly difficult to resolve include the sister-group relationship of Ceratopogonidae + Chironomidae, which was recovered only when 1150 genes were analysed (under ML), with only 72% of RADICAL trees supporting this relationship, mirroring the low BS support of this clade. The FcLM results (Fig. 4), showed support for the sister-group relationship between Ceratopogonidae and Chironomidae (T3 = 78%), which is congruent with the ML topology inferred from the amino acid dataset. Alternative relationships had poor support: T1 (Chironomidae sister group to the remaining Culicomorpha) = 21%, and T2 (Ceratopogonidae sister group to the remaining Culicomorpha) = 2%. The FcML results suggested that the sister-group relationship between Chironomidae + Ceratopogonidae is indeed better supported by ML.

Maximum likelihood topologies obtained from the nucleotide dataset (Fig. S1) differed in the placement of Dixidae and Podonominae, whereas the ML topologies (Fig. S2) obtained with the reduced nucleotide datasets were congruent with the amino acid based trees. Parsimony-based trees obtained with the full nucleotide dataset yielded unusual results: some outgroups were placed within the ingroup (Fig. S3), while the removal of third positions yielded a topology that conflicted with the ML amino acid tree with regard to the placement of Ceratopogonidae + Chironomidae (T3 = 78%), which is congruent with the ML topology inferred from the amino acid dataset. The FcML results suggested that the sister-group relationship between Chironomidae + Ceratopogonidae + Simulidae is indeed better supported by ML.
Discussion

The monophyly of Culicomorpha is strongly confirmed by parsimony and ML analyses (BS, JK = 100) (Fig. 3). Within Culicomorpha, most relationships are robust and congruent across analyses based on the amino acid dataset. Some unexpected results obtained from nucleotide datasets inclusive of third codon positions mirror similar results obtained in recent transcriptome-based phylogenomic studies (Breinholt & Kawahara, 2013; Misof et al., 2014). Overall, we find that all strongly supported nodes are congruent between topologies from the reduced nucleotide datasets and amino acid dataset and we consider the amino acid topology our best estimate of phylogenetic relationships. The superfamily Culicoidea is monophyletic (BS, JK = 100) and Corethrellidae (Chaoboridae + Culicidae) – the core Culicoidea – is fully supported (BS, JK = 100). Dixidae forms the sister group to this group (BS = 100, JK = 92). All families that are represented by two or more taxa are strongly supported as monophyletic; namely Dixidae (two taxa), Corethrellidae (two taxa) and Culicidae (seven taxa). Within Culicidae, the monophyletic subfamily Anophelinae is sister to Culicininae. This subfamily is estimated as (Aedini + Culicinae) (Toxorhynchitini + Sabethini), with all nodes having full support (BS, JK = 100).

In contrast, Chironomoidea is a grade subtending Culicoidea. A robust clade comprising Simuliidae + Thaumaleidae (BS, JK = 100) is sister to Culicoidea (BS, JK = 100). Chironomidae and Ceratopogonidae are recovered as sister groups albeit with weak support in the likelihood analysis (BS = 71), while Ceratopogonidae is placed as sister group to remaining Culicomorpha with strong support on the maximum parsimony tree (MPT) (JK = 100). The FcLM suggests that the sister-group relationship between Chironomidae and Ceratopogonidae is indeed better supported by ML. Three families are sampled sufficiently to discuss internal relationships: Culicidae, Chironomidae and Ceratopogonidae. Within Culicidae, the subfamily relationships are compatible with Harbach & Kitching (1998) while tribal relationships are compatible with the multi-gene Bayesian analyses of Reidenbach et al. (2009), namely Anophelinae sister to Culicinae, but differ in that the representative of Culicinae (Culex) is sister only to the species of Aedes (tribe Aedini) rather
than all other Culiciniae sampled, i.e. (Aedini + Culicini) (Toxorhynchitini + Sabethini); all nodes are strongly supported. Within Chironomidae, the multiply sampled subfamilies Tanypodinae and Diamesinae, Prodiamesinae, Orthocladiinae are monophyletic, and the relationships within this family are robust and overall follow those recovered by Cranston et al. (2012) based on a Bayesian analysis of four genes. Our results only differ in the placement of the subfamily Telmatogotoninae which is sister group to Diamesinae alone rather than to a wider clade comprising Diamesinae + (Prodiamesinae + (Orthocladiinae + Chironominae)), as suggested by Cranston et al. (2012). Although Ceratopogoninae is represented in our sampling by only three subfamilies, the relationship Ceratopogoninae + (Dasyheleinae + Forcipomyiinae) is as proposed based on analysis of the mitochondrial gene COI (Beckenbach & Borkent, 2003) and on morphology (Borkent, 2012). Improved taxon sampling within families will help understand wider relationships.

Although the vastly expanded data derived from transcriptomes provide resolution and high support for most nodes, some clades remain poorly or ambiguously supported. Our ML tree (Fig. 3) recovers Edward’s original hypothesis of a sister-group relationship between Ceratopogonidae and Chironomidae (‘midges’; e.g. Edwards 1929). The clade Ceratopogonidae + Chironomidae is sister to the remaining Culicomorpha, but only with modest support. However, in the parsimony analysis (Fig. 3), the basal split of Culicomorpha is a bifurcation consisting of Ceratopogonidae and the remaining Culicomorpha. The RADICAL likelihood analysis revealed that more than 93% of the genes were required to fix this node with only 72% of the RADICAL likelihood trees supporting this clade, implying that the signal per partition (gene) is weak. The FcLM mapping supports Ceratopogonidae + Chironomidae (T3) on only 78% of all drawn quartets, with the alternate T2 topology (Ceratopogonidae sister to all other Culicomorpha) having a support of 21%. The RADICAL analysis as well as FcLM suggests that the available dataset cannot resolve the placement of Ceratopogonidae with confidence.

In contrast to this poorly supported unstable relationship at the base, many previously poorly supported nodes in molecular analyses (e.g. Wiegmann et al., 2011) now find better support based on the more than 1000 loci derived from transcriptomic data. In particular, the Dixidae, if included in Culicoidea, finds strong support as sister group of the remaining Culicoidea in the ML tree, with only the parsimony analysis yielding lower support (Fig. 3). Several previous studies had found conflicting positions for Thaumaleidae (see Fig. 2), but it finds strong support as sister family to Simuliiidae in our study. However, Simuliiidea (sensu Borkent, 2012) is rejected because Ceratopogonidae is not supported as sister group of Thaumaleidae + Simuliiidae in either the ML or parsimony analyses. Overall, our results suggest that the value of ‘superfamilies’ within Culicomorpha is questionable because only Culicoidea is demonstrably monophyletic, whereas Chironomoidea is clearly a paraphyletic grade.

Fortunately, weak support for the basal split within Culicomorpha matters little when discussing competing hypotheses regarding the function of mouthparts (and therefore potential blood-feeding) in these flies. Although many authors have discussed gains and losses of mouthparts modified for blood-feeding, only Borkent (2012) provides a detailed discussion in an explicit phylogenetic context. Borkent’s tree (see Fig. 2G) differs from ours only with regard to support and relationships at the base (position of Chironomidae), so that his overall conclusions are identical. Adult female biting mouthparts belong to the groundplan of the Culicomorpha ancestor. Depending on the as-yet-unknown sister group, such mouthparts may be plesiomorphic for Diptera. Each family of Culicomorpha that lacks biting mouthparts (Thaumaleidae, Dixidae and Chaoboridae) is sister to a biting group. With regard to toothed mandibles in ‘nonbiting’ midges (Chironomidae), the generally accepted hypotheses remain valid: functional mandibles are present only in southern African Archaeochlus Brundin and Australian Austrochlus Cranston (subfamily Podonominae) amongst extant taxa but are widely distributed in amber fossils, including some Tanypodinae, Podonominae and the Aenemeinae (Cranston et al., 2012), the putative sister to all extant taxa. It remains to be tested via observations of adult flies in the field whether the feeding mode of Archaeochlus/Austrochlus is similar to certain insect-feeding Ceratopogonidae (Borkent, 2012), as suggested by mandible morphology. Losses of the mandibles are common in most families, including Corethrellidae, Culicidae, Simuliidae, Ceratopogonidae and Chironomidae.

The value of the more than a 100-fold expansion of gene coverage for entomological phylogenomic studies is evident (e.g. Niehuis et al., 2012; Letsch & Simon, 2013; Misof et al., 2014; Kang et al., 2017; Bazinet et al., 2017; Peters et al., 2017). For example, the once controversial monophyly of Polyneoptera and positions therein of Strepsiptera as sister to Coleoptera and Hymenoptera as sister group to ‘Aparaglossata’ have become widely accepted. Nevertheless, resolution with support for some relationships, such as the monophyly of Palaeoptera and Paraneoptera, remain intractable. Furthermore, phylogenomics can generate strong support for nodes that are in conflict depending on analysis method, i.e. statistical support should be handled cautiously as large data can only account for statistical but not for systematic bias in data and methods (Tarrio et al., 2001; Li et al., 2014; Kück & Wägele, 2016).

There is little doubt that taxon sampling remains critical for addressing phylogenetic relationships that are difficult to resolve (Bazinet et al., 2017) even when thousands of genes can be sampled for a small number of taxa. Low/ambiguous support for our basal node within Culicomorpha may be due to sampling, e.g. the Ceratopogonidae is represented by three taxa only and it is unfortunate that the Leptoconopinae (especially the Cretaceous and extant genus Austroconops Wirth & Lee) was unsampled. Sampling in the Chironomidae is denser, but important taxa such as Buchonomyiinae, which has been suggested as the putative sister group to all extant Chironomidae (Cranston et al., 2012), is lacking, as is the early branching Aphroteniinae. Note, however, that in failing to resolve the basal node of Culicomorpha with confidence despite millions of nucleotides, we are not alone as
noted earlier with several well-known and near intractable nodes in Insecta (Misof et al., 2014).

Other practical issues arise in transcriptomic studies. One pertains to obtaining and management of vouchers. Specimens must be collected alive into RNAlater or flash-frozen. Although the amount of tissue needed has diminished over the course of our project, most individuals of the smallest insects, such as nearly all Ceratopogonidae, Aphroteniinae and Podonominae (Chironomidae), and indeed many other Chironomoidea provide inadequate quantities of RNA unless the RNAs of multiple individuals were combined. This risks accidental multispecies sampling for taxa, especially if individuals are identifiable only with slide preparation and microscopy as discussed by Stur & Borkent (2014) in relation to barcoding of Ceratopogonidae. Rare taxa are usually only available as single individuals and it is tempting to extract from the entire specimen, thereby sacrificing the ability to validate the identity with morphology. In our study, these issues were addressed either by vouchering important morphological structures for identification purposes or through rearing and/or DNA barcoding. Larvae of small culicomorphans were reared from pitcher plants and the RNA was extracted individually. We then used DNA barcoding to verify that multiple specimens belonged to the same species belonging to the desirable group before pooling RNA extractions. For one species, we eventually had to pool the RNA for three specimens (T. aranoides) in order to have sufficient RNA quantity. We were less careful with regard to one sample of what had been identified as larval thaumaleids in the field. A contamination with athericerids was only detected when the taxon was placed outside of Culicomorpha in preliminary analyses. The misidentification was retrospectively discovered via DNA barcoding (COI). This detection encouraged a renewed search for a correct thaumaleid, a critical taxon to Culicomorpha phylogenetics, but due to the lack of a voucher this error could have easily perpetuated the finding in the literature that this group is an irreconcilable ‘misfit’. Public databases are replete with such problems, and the inability to obtain morphological vouchers for subsequent verification is evident from Webb et al. (2012) and explicit in Stur & Borkent (2014).

Finally it is instructive to ask how several trees illustrated in Fig. 2 came so close to the result that is supported here based on thousands of genes. Some morphological studies drawn from all life stages, as in Wood & Borkent (1989) and Oosterbroek & Courtney (1995) (Fig. 1B), provided a closer estimate than did that of Sæther (2000). Cranston et al. (2010) argued that congruence in molecular and morphological analyses in Podonominae was because morphological matrices in Culicomorpha derived from more extensive data (larva, pupa, adults of both sexes) than is typical in systematic research in other Holometabola. This approach apparently worked well for Culicoida, with multiple morphological analyses yielding similar results. However, this was less successful for chironomoidan grade in which morphological and molecular data compete to place some taxa with confidence.

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: 10.1111/syen.12285

Table S1. List of taxa obtained from public databases, sources, number of contigs (average coverage > 10, length > 100 bp), number of genes from orthology prediction.

Table S2. Reference data for orthologue set.

Table S3. FeLM results.

Fig. S1. Maximum likelihood topology of nucleotide dataset with bootstrap support values

Fig. S2. Maximum likelihood topology of reduced nucleotide datasets with bootstrap support values

Fig. S3. Maximum parsimony topology of nucleotide dataset with jackknife support values

Fig. S4. Maximum parsimony topology of reduced nucleotide datasets (third codon postions removed) with jackknife support values

Fig. S5. Maximum parsimony topology of reduced nucleotide datasets (only second codon postions) with jackknife support values

Fig. S6. Heat map showing species-pairwise amino acid site coverage for the dataset from Allstat. Shared site coverage is shown from shades of blue (low) to white (high).

Fig. S7. Maximum likelihood topology of amino acid dataset before mare reduction with bootstrap support values

Acknowledgements

We thank especially our dipterist colleagues Art Borkent, Brad Sinclair, Ralph Harbach and John (Kevin) Moulton for interest, advice and discussions on taxonomy and identifications. Steve Marshall generously allowed us to modify his excellent photographs in Fig. 1, from Flies: The Natural History and Diversity of Diptera published by Firefly Books Ltd, Buffalo, New York (2012). Matt Krosch of Queensland University of Technology, Brisbane, reviewed a draft of this manuscript and Art Borkent provided perspicacious insights. We acknowledge grant support from NUS SEABIG (R-154-000-648-646 and R-509154-000-648-733). The authors declare no conflict of interest.

References


Niehuis, O., Hartig, G., Grath, S. *et al.* (2012) Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. *Current Biology*, **22**.


Accepted 9 January 2018