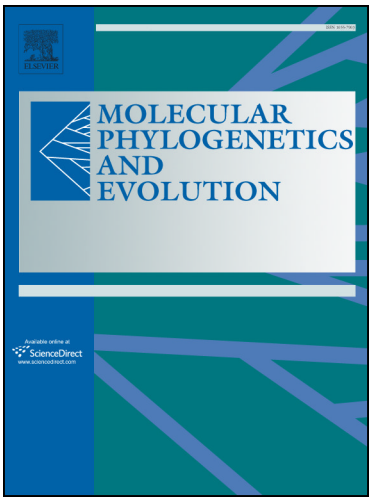

Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids

Jessica P. Gillung, Shaun L. Winterton, Keith M. Bayless, Ziad Khouri, Marek L. Borowiec, David Yeates, Lynn S. Kimsey, Bernhard Misof, Seunggwan Shin, Xin Zhou, Christoph Mayer, Malte Petersen, Brian M. Wiegmann

PII: S1055-7903(18)30223-9
DOI: https://doi.org/10.1016/j.ympev.2018.08.007
Reference: YMPEV 6254

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 5 April 2018
Revised Date: 3 August 2018
Accepted Date: 7 August 2018

Please cite this article as: Gillung, J.P., Winterton, S.L., Bayless, K.M., Khouri, Z., Borowiec, M.L., Yeates, D., Kimsey, L.S., Misof, B., Shin, S., Zhou, X., Mayer, C., Petersen, M., Wiegmann, B.M., Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids, *Molecular Phylogenetics and Evolution* (2018), doi: https://doi.org/10.1016/j.ympev.2018.08.007

Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids

Jessica P. Gillung[a,b,*], Shaun L. Winterton[b], Keith M. Bayless[c], Ziad Khouri[a], Marek L. Borowiec[d], David Yeates[e], Lynn S. Kimsey[a], Bernhard Misof[f], Seunggwan Shin[g], Xin Zhou[h], Christoph Mayer[f], Malte Petersen[f], Brian M. Wiegmann[i]

[a]Bohart Museum of Entomology, University of California, One Shields Ave., Davis, CA 95616, USA
[b]California State Collection of Arthropods, 3294 Meadowview Rd, Sacramento, CA 95832, USA
[c]California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA
[d]School of Life Sciences, Social Insect Research Group, Arizona State University, Tempe, AZ 85287, USA
[e]National Research Collections Australia, Clunies Ross Street, Acton, ACT 2601**,** GPO Box 1700, Canberra, ACT 2601, Australia
[f]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany
[g]Department of Biological Sciences, University of Memphis, 3700 Walker Avenue, Memphis, TN 38152, USA
[h]Department of Entomology, China Agricultural University, Beijing, China 100193
[i]Department of Entomology & Plant Pathology, North Carolina State University, 3114 Gardner Hall, Raleigh, NC 27695-7613, USA

* Corresponding author.
Email address: jpgillung@ucdavis.edu (J.P. Gillung).

Abstract

The onset of phylogenomics has contributed to the resolution of numerous challenging evolutionary questions while offering new perspectives regarding biodiversity. However, in some instances, analyses of large genomic datasets can also result in conflicting estimates of phylogeny. Here, we present the first phylogenomic scale study of a dipteran parasitoid family, built upon anchored hybrid enrichment and transcriptomic data of 240 loci of 43 ingroup acrocerid taxa. A new hypothesis for the timing of spider fly evolution is proposed, wielding recent advances in divergence time dating, including the fossilized birth-death process to show that the origin of Acroceridae is younger than previously proposed. To test the robustness of our phylogenetic inferences, we analyzed our datasets using different phylogenetic estimation criteria, including supermatrix and coalescent-based approaches, maximum-likelihood and Bayesian methods, combined with other approaches such as permutations of the data, homogeneous versus heterogeneous models, and alternative data and taxon sets. Resulting topologies based on amino acids and nucleotides are both strongly supported but critically discordant, primarily in terms of the monophyly of Panopinae. Conflict was not resolved by controlling for compositional heterogeneity and saturation in third codon positions, which highlights the need for a better understanding of how different biases affect different data sources. In our study, results based on nucleotides were both more robust to alterations of the data and different analytical methods and more compatible with our current understanding of acrocerid morphology and patterns of host usage.

Key words: Bayesian inference; bioinformatics; conflict; Diptera; fossilized birth-death process; systematic error.

1. Introduction

The size of molecular datasets in phylogenetics has been growing greatly since the introduction of high-throughput sequencing. The combination of the advances in genomic data acquisition with new bioinformatics tools has resulted in a novel field of evolutionary biology, phylogenomics. The onset of phylogenomics has resolved some of the most challenging evolutionary questions while giving us a new perspective on biodiversity (e.g., Misof et al., 2014; Prum et al., 2015; Garrison et al., 2016; Hamilton et al., 2016; Kocot et al., 2016; Branstetter et al., 2017; Shin et al., 2017; Espeland et al., 2018; Winterton et al., 2018).

Increasing the quantity of phylogenomic data successfully alleviates stochastic error caused by limited data sampling, but the impact of systematic error is potentially augmented (Yeates et al., 2016). Several sources of systematic error have been identified, including compositional heterogeneity, missing data, heterogeneity in evolutionary rates among lineages, among others (Felsenstein, 1978; Jermiin et al., 2004; Bininda-Emonds 2007; Lartillot et al., 2007; Edwards 2009; Nabholz et al., 2011; Roure et al., 2013; Mirarab et al., 2014; Goremykin et al., 2015; Streicher et al., 2016). Thus, merely increasing the number of gene sequences in datasets does not necessarily resolve all phylogenetic incongruence. Instead, a number of cases have been observed in which alternative phylogenomic datasets strongly support conflicting conclusions, each with highly resolved phylogenetic estimates and maximal nodal support values (e.g., Crawford et al., 2012; Shaffer et al., 2013; Wang et al., 2013; Jarvis et al., 2014; Chang et al., 2015; Pisani et al., 2015; Prum et al., 2015). Phylogenetic conflict, however, can originate not only from different datasets, but also from alternative coding of the same data (Fučíková et al., 2016). Protein-coding genes can be analyzed as amino acids, nucleotides or codons, and choosing which data type to analyze in phylogenomics is a challenge that could significantly affect reliability and confidence of the results.

In the case of phylogenetic studies that focus on recent divergences, nucleotides are probably more informative than amino acids. This is because substitutions are more likely to have occurred at synonymous sites. For deep divergences, however, the choice is not as straightforward. Even though analyses of amino acid datasets are suggested to be less prone to systematic error due to compositional heterogeneity across sites and taxa (Jeffroy et al., 2006; Rodríguez-Ezpeleta et al., 2007; Rota-Stabelli et al., 2013), the statistical phylogenetic analysis of amino acid data presents challenges beyond those often faced with the analysis of DNA sequences. Most approaches to the analysis of amino acid datasets make use of empirical amino acid models, in which all of the potentially free parameters are fixed to specific values estimated from a large number of sequences (Dayhoff et al., 1978; Henikoff and Henikoff 1992; Jones et al., 1992; Adachi and Hasegawa 1996; Cao et al., 1998; Adachi et al., 2000; Whelan and

Goldman 2001; Le et al., 2012). Although the fixed amino acid models succeed in reducing the number of free parameters to be estimated, it is possible that even the best-fitting fixed amino acid model is not particularly appropriate for the data at hand. Consequently, if the model is misspecified, the phylogeny estimate might be inaccurate, potentially resulting in conflicting estimates of phylogeny under either nucleotides or amino acids. Conflict among topologies due to alternative data coding as nucleotides or amino acids is relatively common in phylogenomics (Zwick et al., 2012; Rota-Stabelli et al., 2013; Cox et al., 2014; Reddy et al., 2017; Shin et al., 2017; Haddad et al., 2018), but our knowledge of systematic error in big-data phylogenetics is still incipient.

Here, we attempt to understand the basis for the incongruence among phylogenomic trees originating from alternative data types (nucleotides versus amino acids) by investigating the evolution of spider flies (Acroceridae), the only family of flies that exclusively parasitize spiders. Acroceridae is a relatively ancient and morphologically derived lineage of lower Brachycera, consisting of a charismatic and remarkably diverse assemblage of insects. Spider fly origins have been estimated in the Early Mesozoic (173–221 MYA) (Winterton et al., 2007), but their fossil record extends only to the Upper Jurassic (~150 MYA) (Gillung and Winterton 2017). Species of Acroceridae attack spiders in 26 families (Cady et al., 1993; Gillung and Borkent 2017) and are currently distributed in 55 genera and approximately 530 species (Winterton et al., 2007; Schlinger et al., 2013). Three subfamilies are recognized, Acrocerinae, Panopinae and Philopotinae. Monophyly of Philopotinae is based on a series of morphological characters, while Panopinae is defined based on their unique mygalomorph attacking behaviour. The monophyly of Acrocerinae is contentious, and its internal relationships are poorly known (Winterton et al., 2007). Thus, additional data and analyses are needed to test the monophyly of the subfamilies and to establish a robust higher-level classification.

In this study, we address the two-fold problem of data type choice and Acroceridae relationships, bringing the greatly expanded gene sampling of anchored phylogenomics to bear on spider fly phylogeny. We recovered 240 unique orthologous loci of 43 species representing all major lineages of spider flies, plus seven representatives of outgroup families. Through the integration of high-throughput sequencing and comparative methods, we provide a robust hypothesis for the pattern and timing of spider fly evolution. Using Acroceridae as a system, we explore the potential of genomic data to resolve relationships in relatively ancient radiations and explore the effects of potential confounding factors in phylogenomic reconstruction.

2. Material and Methods

2.1. Taxon sampling and DNA acquisition

Taxa were carefully selected to represent the greatest diversity within Acroceridae and to ensure as close to proportional sampling as possible, based on ongoing taxonomic studies (Gillung and Winterton 2011; Winterton and Gillung 2011; Schlinger et al., 2013; Borkent et al., 2016; Gillung and Nihei 2016). Newly generated Anchored Hybrid Enrichment (AHE) data for 42 species of Acroceridae plus the transcriptome of one additional spider fly species were included as the ingroup. Transcriptomes of six species and AHE data of one species in the lower Brachycera were used as outgroup taxa, representing the families Asilidae, Bombyliidae, Hilarimorphidae, Nemestrinidae, Pantophthalmidae, Tabanidae and Xylophagidae (Supplementary Table 4). Genetic material was extracted from the legs and thorax, with genitalia, remaining legs, head and wings preserved in 95% ethanol as vouchers (Supplementary Table 4). DNA was extracted from frozen specimens preserved in 95% ethanol using a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). RNA was extracted from specimens preserved in RNAlater following methods described in Misof et al. (2014) and Peters et al. (2017). AHE capture was carried out following the general methods of Lemmon et al. (2012) for sonication, library preparation, indexing and enrichment. Probes were developed specifically for Diptera at the Center for Anchored Phylogenomics at Florida State University, as described in Young et al. (2016). The AHE Diptera Probe Set targets 559 loci, with sequences publicly available as supplementary information from Young et al. (2016). AHE data was sequenced as single reads, with up to 48 multiplexed samples per lane on an Illumina MiSeq platform at the NCSU Genomic Sciences Laboratory (Raleigh, NC). Transcriptome libraries were prepared following methods described by Misof et al. (2014) and Peters et al. (2017). Reads were inspected for quality with Fastqc (Andrews 2010) and trimmed with Trimmomatic (Bolger et al., 2014), with minimum per base sequence quality set to 20, and minimum read length set to 25 bp.

2.2. Sequence assembly and orthology prediction

De novo assemblies were carried out using Trinity v. 2.2 (Grabherr et al., 2011). For data provided by 1KITE, raw reads were quality checked, assembled with SOAPdenovo-Trans-31kmer (version 1.01) (Xie et al., 2014) and cleaned from potential contaminants as described by Peters et al. (2017). We used Orthograph v.0.5.8 (Petersen et al., 2017) to infer orthology of sequence contigs, with single copy genes extracted and assembled from OrthoDB5 (Waterhouse et al., 2011) and reciprocal search set to relaxed. Orthologous genes were identified based on an ortholog reference set of 3,288 orthologous clusters of sequences groups (single copy protein-coding genes) from five reference species: *Anopheles gambiae* Giles, *Tribolium castaneum* (Herbst), *Drosophila melanogaster* Meigen, *Mayetiola destructor* (Say) and *Bombyx mori* (Linnaeus) (Kutty et al., 2018). Following orthology prediction,

contaminant viral, bacterial and fungal sequences were identified using NCBI BLAST; loci not matching Diptera or other insects were removed.

2.3. Dataset construction

Internal stop codons and "U" (Selenocysteine) were replaced with an "X" in the amino acid dataset and with "NNN" on the nucleotide dataset, respectively. Amino acid sequences were aligned using MAFFT v.7.123b (Katoh and Standley 2013) with the *L-INS-i* algorithm. Ambiguously or randomly aligned sections identified by Aliscore v2.2 (Misof and Misof 2009; Kück et al., 2010) were removed from the amino acid alignment, and the corresponding codons from the nucleotide loci were removed using Alicut and custom Perl scripts (Misof and Misof 2009; Kück et al., 2010). Nucleotide sequences were then aligned using the amino acid alignment as blueprint in Pal2Nal (Suyama et al., 2006), using a slightly modified version (see Misof et al., 2014). Individual loci were concatenated using AMAS (Borowiec 2016). We combined transcriptomic and AHE data for all 50 species included in this study, which resulted in a dataset containing 3,234 genes. Because many of these genes were present only in the taxa represented by transcriptomes, which leads to non-random distribution of missing data as these were mainly outgroups, we filtered loci based on taxon occupancy, keeping only the loci present in at least 24 taxa (out of 50). The final nucleotide and amino acid datasets contained 240 loci, with 172,905 base pairs and 57,635 amino acid sites, respectively (Supplementary Files 1–4).

2.4. Dataset exploration

Pairwise sequence comparisons using Bowker's matched-pairs tests of symmetry (Bowker 1948) were performed in SymTest version 2.0.47 (https://github.com/ottmi/symtest) (Jermiin et al., 2004; Ababneh et al., 2006). The software was also used to generate heat maps based on the inferred p-values, using default window and step sizes. We applied Bowker's test as implemented in SymTest on the amino acid dataset, and on the nucleotide dataset with and without 3rd codon positions.

2.5. Phylogenetic Analyses

Both supermatrix and species tree approaches were used for tree estimation on both amino acid and nucleotide datasets. We performed multiple alternative rooting strategies to account for uncertainty in the placement of Acroceridae within the lower Brachycera (e.g., Wiegmann et al., 2011; Shin et al., 2018). Alternative rooting along branches of the outgroup did not affect the relationships within

Acroceridae (results not shown), thus we arbitrarily constrained Pantophthalmidae as the root in the topologies presented here. Multispecies coalescent analyses (MSC) were performed using ASTRAL v4.9.7 (Mirarab and Warnow 2015), with gene trees estimated using RAxML v.8.2.10 (Stamatakis 2014), and branch support values calculated using 500 bootstrap replicates from RAxML. For the concatenated analyses, alignments were initially partitioned by genes, which were then grouped into meta-partitions using PartitionFinder 2 (Lanfear et al., 2016), with the *rcluster* search algorithm (Lanfear et al., 2014) and BIC for model selection. For the nucleotide model selection analyses, we did not include the GTR+I+G mixture model because this approach has been demonstrated to result in undesirable interactions among parameters (Yang 1993, 1996, 2006; Sullivan et al., 1999; Mayrose et al., 2005; Jia et al., 2014). Model selection for the amino acid dataset was performed including all models available in PartitionFinder 2, using the *--raxml* option. The best fitting model was selected using BIC. Basic alignment statistics, including percentage of missing data, A/T and G/C content, alignment length and proportion of variable sites were obtained using AMAS (Borowiec 2016). ExaML (Kozlov et al., 2015) was used to estimate phylogenies under Maximum Likelihood (ML), with parsimony starting trees inferred with RaxML v8.2.10. Node support was estimated via slow non-parametric bootstrapping, with 500 bootstrap replicates per dataset generated with RaxML. Ten different ExaML tree searches were performed and compared with each other to ensure that the analyses were not trapped in a local optimum (i.e., the same topology was recovered). Bayesian tree inference (BI) was carried out by running four independent replicates, with four chains each, using either ExaBayes v1.4 (Aberer et al., 2014) or MrBayes (Ronquist and Huelsenbeck 2003) through the Cipres Science Gateway v3.3 (Miller et al., 2010). Runs were carried on for at least 50,000,000 generations and were sampled every 1,000 generations. Branch lengths were linked among partitions and a relative burn-in of 25% was used. Convergence was evaluated by ensuring effective sample size values (ESS) greater than 200 for each parameter in Tracer v1.6 (Rambaut et al., 2014), as well as potential scale reduction factors (PSRF) ranging close to one and average standard deviations of split frequencies (ASDSF) smaller than 0.01%.

A site-heterogeneous CAT-GTR-G mixture model (Lartillot and Philippe 2004) was implemented in PhyloBayes (Lartillot et al., 2009). Two independent Markov chains with a total length of 10,000 cycles were run for each analysis, with the first 4,000 trees being discarded as burn-in and the posterior consensus determined using the remaining 6,000 trees. Convergence between the two chains was assured, with the largest discrepancy observed across all bipartitions (maxdiff) being less than 0.1.

We implemented the *degen1* v1.4 approach (Regier et al 2010; Zwick et al., 2012) to mask synonymous signal and keep only non-synonymous changes at all coding positions. The degenerated alignment was initially partitioned by locus, and the best fitting substitution model and partition scheme were selected using PartitionFinder 2 as described above. Analysis of the degenerated nucleotide dataset

was executed in MrBayes via the Cipres Science Gateway, with four coupled chains and settings as described previously.

2.6. Substitution rate heterogeneity

We used a simplified binning approach as proposed by Mirarab et al. (2014), but grouping genes based on rate of evolution as opposed to bootstrap values on branches as originally proposed by the authors. Gene trees were estimated under ML in RaxML v8.2.10, using the best-fitting model identified by PartitionFinder 2 for each locus both as amino acids and nucleotides. Utilizing the gene_stats R script used in Borowiec et al. (2015), we inferred the average branch lengths, used here as a proxy for rate of evolution, with short branch lengths indicating relatively slowly evolving loci, and long branch lengths indicating relatively faster evolving loci. After sorting genes based on average branch lengths (lowest to highest), we divided the entire set of 240 loci – both as amino acids and nucleotides – into three subsets of 80 loci, so that each subset consisted of a set of loci evolving at roughly under the same rate – namely 'slow', 'intermediate' and 'fast'. We discarded the intermediate population of average branch lengths to ensure two discrete loci populations separated by a large buffer population and concatenated the genes in each of the fast and slow subsets. We then estimated phylogenetic trees separately for each subset using BI in MrBayes 3.2 via the Cipres Science Gateway as described above and assessed their topological congruence with the tree generated from all loci.

2.7. Four-cluster likelihood mapping

We performed four-cluster likelihood mapping (FcLM; Strimmer and von Haeseler 1997) to quantify the support for the monophyly of Panopinae in the amino acid and nucleotide datasets as implemented in IQTree (Nguyen et al., 2015). We defined four taxon clusters: Panopinae1 (7 species), Panopinae2 (7 species), *Turbopsebius* Schlinger + *Cyrtus* Latreille (2 species), and *Psilodera* Gray + *Pterodontia* Gray (4 species). All remaining species were ignored during analyses. IQTree analyses were conducted using the -m TEST option, which implements ModelFinder (Kalyaanamoorthy et al., 2017) to automatically select the best fitting model, with alignment partitioned by locus.

2.8. Divergence times estimation

The chronogram for Acroceridae was estimated using BEAST v2.4.7 (Bouckaert et al., 2014). We used the nucleotide alignment for the dating analyses and removed six outgroups to enforce proportional

sampling of terminal taxa, keeping only *Hilarimorpha* Schiner (Hilarimorphidae) as outgroup. Because the whole nucleotide alignment is too large for a computationally feasible BEAST analysis, we used a method of matrix reduction as implemented in MARE (Misof et al., 2013) using phylogenetic information content as the criterion for keeping or removing genes from the analysis. We applied MARE on the amino acid dataset using the default settings and then reduced the nucleotide dataset accordingly. MARE reduced the nucleotide dataset from 240 to 65 genes, increasing the overall information content of the alignment from 0.31 to 0.45. We applied PartitionFinder 2 using linked branch lengths, the rcluster algorithm and BIC to select the statistically best-fit partitioning scheme and models of nucleotide substitution available in BEAST 2. We used an uncorrelated relaxed molecular clock model (Drummond et al., 2006) and a lognormal prior, with tree and clock model linked across partitions. Fossils included as terminals in the FBD analyses are provided in Supplementary Table 3. Because we did not include morphological data in our analysis to place the fossils in a "total evidence" dating framework *sensu* Ronquist et al. (2012), we assigned them to appropriate groups via monophyly constraints (Heath et al., 2014) according to a recent review of spider fly fossils by Gillung and Winterton (2017). The two Jurassic species of *Archocyrtus* were treated as stem acrocerids, while the Cretaceous-aged *Schlingeromyia minuta* was included within the crown Acroceridae. *Glaesoncodes completinervis* was treated as stem *Ogcodes* based on head and wing venation characters, while *Ogcodes exotica* was included in the crown *Ogcodes*. Finally, *Cyrtinella flavinigra* and *Villalites electrica* were placed in a clade containing *Cyrtus* and *Turbopsebius* also based on head and wing venation characters (Gillung & Winterton 2017).

We ran the analysis for over 600 million generations with four incrementally heated chains and evaluated MCMC convergence and mixing in Tracer v1.6, ensuring that effective sample sizes (ESS) exceeded 200 for all parameters. We then resampled the phylogenetic trees at a lower frequency in LogCombiner v2.3.1 (BEAST package), with a burn-in of 30%. Finally, we summarized the subsampled trees in a maximum clade credibility tree using TreeAnnotator v2.3.1 (BEAST package), with mean heights as node heights. We further compared the effective prior (under the prior) and posterior distributions (with data included) of all the parameters to ensure that our analyses were not prior-sensitive and that the data were informative for the MCMC analyses (results not shown).

2.9. Data availability

Published AHE and transcriptome raw data for 44 species included herein is available from the NCBI SRA database (Bioprojects PRJNA325838). Transcriptome raw reads for the remaining six species will be available in the near future according to the 1KITE Project timeline (http://www.1kite.org/). Accession numbers for the published data and unique identifiers for the 1KITE unpublished

transcriptomic data used here are provided in Supplementary Table 4. Individual loci used in this study can be obtained from the alignment files (Supplementary Files 1–4) in the Zenodo Database (https://doi.org/10.5281/zenodo.1289998) prior to the release of 1KITE transcriptomic raw data.

3. Results

3.1. Incongruence of nucleotide and amino acid-based phylogenies

The analyses of the nucleotide and amino acid datasets under a variety of tree estimation methods and dataset permutations resulted in two well-supported topologies, one based on amino acids and the other based on nucleotides (Fig. 1). The phylogeny based on nucleotides was well supported throughout, with a Bayesian posterior probability (PP) of 1 on each node, and only three nodes with maximum likelihood bootstrap values (BS) lower than 100% (Fig. 1A). The phylogeny based on the concatenated amino acid alignment was less supported overall, with 25% of nodes with PP and BS lower than 1.0 and 100%, respectively, and some of the poorly supported nodes located along the backbone of the tree (Fig. 1B). The multispecies coalescent (MSC) analysis using nucleotides resulted in a topology very similar to the one based on concatenation and was relatively well supported overall (Supplementary Fig. 1). The MSC topology based on amino acids (Supplementary Fig. 1) was congruent with the one based on the concatenated dataset (Fig. 1B), albeit with weak statistical support, with many of the particularly poorly supported nodes placed along the backbone.

The monophyly of Panopinae and its internal relationships represent the most significant difference between the phylogenies based on amino acids and nucleotides. In the topology based on nucleotides, Panopinae was recovered as monophyletic and sister to a clade including representatives of the former Acrocerinae (Fig. 1A). In contrast, the topology based on amino acids recovered a polyphyletic Panopinae, with two lineages formerly included in Acrocerinae nested within the subfamily (Fig. 1B). Establishing the evolutionary history of Panopinae has profound implications in the understanding of Acroceridae host usage. Species of Panopinae are unique among acrocerids in attacking heavy bodied, stout legged spiders in the Mygalomorphae, including tarantulas, trapdoor spiders, funnel-web spiders, among others. All other spider flies attack hosts in the Araneomorphae, such as jumping spiders, wolf spiders, orb-weavers, among many others (Gillung and Borkent 2017). Assuming that Panopinae is monophyletic would result in a hypothesis for Acroceridae evolution where there was only one invasion of Mygalomorphae, while the alternative hypothesis of non-monophyly would require either two independent origins for the mygalomorph host life history, or the loss of this trait in some lineages (Fig. 1).

3.2. Exploratory analyses

We constructed nine supplementary datasets and used a plethora of additional analyses to explore the origins of the conflict between nucleotides and amino acids, and to indirectly access the reliability of the two alternative topologies. We removed 3rd codon positions to investigate whether heterogeneity in evolutionary rates across codon positions caused any error in our tree estimation based on nucleotides. The topology based on 1st and 2nd positions only was very similar to the one based on the whole nucleotide dataset, with minor differences in the relationships within some genera (Supplementary Fig. 2). Since there were no significant changes in relationships after the removal of 3rd codon positions, we included all codon positions in downstream analysis of nucleotide data to include as much information as possible.

Additionally, we used a CAT-GTR-G mixture model of base substitution (Lartillot and Philippe 2004), which resulted in a topology that was highly discordant with the nucleotide topology under the homogeneous GTR+G model (Supplementary Fig. 3). The most striking difference was the position of *Ogcodes*, which was recovered as the sister group to Philopotinae under the CAT-GTR-G model (Supplementary Fig. 3). The analysis of amino acid data under the CAT-GTR-G mixture model resulted in a topology that is discordant with all other topologies we recovered based on either nucleotides or amino acids (Supplementary Fig. 3). Similar to the nucleotide topology using the mixture model, *Ogcodes* Latreille was recovered as sister to Philopotinae in the amino acid analysis (Supplementary Fig. 3B).

We also accounted for non-random distribution of missing data by excluding five taxa with low locus coverage. The reduced dataset consisted of 45 taxa (out of 50), which was then analyzed under Bayesian inference. The reduced nucleotide topology (Supplementary Fig. 4) was completely congruent with the topology including all 50 taxa (Fig. 1A), and is well supported overall, with every node having posterior probability (PP) of 1. Conversely, the topology based on the reduced taxon set for amino acids (Supplementary Fig. 4) differs greatly from the topology based on the complete taxon set (Fig. 1B). These results suggest that non-random distribution of missing data had a strong influence on tree estimation using the amino acid dataset, whereas it had no apparent effect on tree estimation based on the nucleotide alignment.

We also investigated the effect of synonymous and non-synonymous information in our analysis. We degenerated nucleotides at codon positions that have the potential to undergo synonymous substitutions using the *degen1* coding approach (Regier et al., 2010; Zwick et al., 2012), and then estimated phylogenies using Bayesian inference. The resulting degenerate nucleotide topology is very similar to the topology based on amino acids, rendering Panopinae polyphyletic, and is relatively well

supported overall, with only five nodes with PP lower than 1 (Supplementary Fig. 5). We also applied the *degen1* coding approach to the nucleotide dataset excluding the five taxa with low gene coverage, thus reducing the effect of non-random distribution of missing data. The resulting degenerate nucleotide topology, in turn, was very similar to the amino acid topology with reduced taxon set, rendering Panopinae paraphyletic (Supplementary Fig. 6).

Moreover, we explored the effects of substitution rate heterogeneity across loci using a simplified binning approach as proposed by Mirarab et al. (2014) (Borowiec 2017; Winterton et al., 2018). We divided the entire set of 240 loci into three subsets of slow-, intermediate- and fast-evolving genes. We then estimated phylogenies under BI separately for the slow- and fast-evolving loci, discarding the intermediate subset of genes. The two topologies based on the nucleotide dataset (for fast- and slow-evolving loci) are completely congruent with each other and highly concordant with the concatenated nucleotide topology (Supplementary Fig. 7). In contrast, the topologies based on the slow- and fast-evolving loci translated as amino acids differ greatly from one another and are highly discordant with the amino acid topology based on all loci (Supplementary Fig. 8). Also, the topologies based on slow and fast-evolving loci as amino acids are poorly supported overall, with some of the low posterior probability (PP) nodes located at the backbone (Supplementary Fig. 8).

To assess the phylogenetic support for the two conflicting hypotheses regarding the monophyly of Panopinae, we implemented a four-cluster analysis with likelihood mapping for the concatenated nucleotide and amino acid datasets, and for each locus separately (Fig. 2). We defined four taxon clusters, two of each containing taxa assigned to Panopinae, and the other two clusters containing taxa of the former Acrocerinae that were recovered nested within Panopinae in the analyses of amino acid data (Fig. 1A; Fig. 2A). Of the three possible unrooted topologies for the four-taxon clusters, only one results in a monophyletic Panopinae (Fig. 2A). Analysis of the concatenated amino acid dataset indicates stronger support for a non-monophyletic Panopinae, with 64.5% of evaluated quartets supporting this hypothesis, while monophyly of Panopinae based on the concatenated amino acid dataset is only supported by 22.7% of the evaluated quartets (Fig. 2C). In contrast, analysis of the concatenated nucleotide dataset indicates stronger support for a monophyletic Panopinae, with 52.6% of all quartets indicating this relationship (Fig. 2C). FcLM analysis of individual loci resulted in a similar scenario. For amino acids, 60.4% of loci supported a non-monophyletic Panopinae, with only 29.2% indicating its monophyly (Fig. 2B, Supplementary Table 1). For nucleotides, alternatively, 46.7% of loci support a monophyletic Panopinae (Fig. 2B, Supplementary Table 1).

We also evaluated whether sequence data in the amino acid and nucleotide datasets (with and without 3rd codon positions) have evolved under globally stationary, time-reversible and homogeneous conditions (SRH) using the software SymTest (Ababneh et al., 2006; Jermiin et al., 2008). Results

indicate that sequences in the nucleotide dataset with 3rd codon positions are unlikely to have evolved under globally SRH conditions, since > 90% of Bowker's tests significantly rejected global symmetry (Fig. 3). The nucleotide dataset without 3rd codon positions, by contrast, suffered much less from such violations, with most pairwise comparisons supporting the hypothesis of homogeneity (Fig. 3). Additionally, results indicate that approximately 10% of sequences in the amino acid dataset are unlikely to have evolved under globally SRH conditions (Fig. 3), with deviations from SRH conditions in the amino acid dataset being much greater than in the nucleotide dataset without 3rd codon positions, but much smaller than in the nucleotide dataset with 3rd codon positions.

3.3. Relationships among Acroceridae lineages

In all resulting phylogenies, Acroceridae was recovered as monophyletic (Fig. 1, Supplementary Fig. 1). The diverse subfamily Acrocerinae was recovered as polyphyletic, consisting of four independent lineages; herein we refer to this non-monophyletic assemblage as the former Acrocerinae. The enigmatic *Carvalhoa* Koçak & Kemal and the cosmopolitan *Acrocera* Meigen were recovered in a clade sister to all other Acroceridae. The genus *Ogcodes* was recovered as sister to the remaining acrocerids (except *Carvalhoa* + *Acrocera*). This relationship was well supported in both amino acid and nucleotide trees regardless of the tree estimation method used. Subsequently, the next clade comprised the subfamilies Philopotinae, Panopinae and two independent lineages of the former Acrocerinae. Within the monophyletic Philopotinae, an early dichotomy was recovered, with *Parahelle* Schlinger and *Thyllis* Erichson in one clade, and *Megalybus* Philippi, *Oligoneura* Bigot and *Philopota* Wiedemann in the other (Fig. 1). The former acrocerine genera *Turbopsebius* and *Cyrtus* were placed in a clade subtending the two remaining lineages, one including the former acrocerine genera *Psilodera* and *Pterodontia*, and the other containing the subfamily Panopinae. Primarily in analyses using amino acids, however, Panopinae was not supported to be monophyletic. Within Panopinae, one basal dichotomy was recovered, with one lineage including *Lasia* Wiedemann, *Eulonchus* Gerstaecker and *Panops* Lamarck, and the other comprising *Pialea* Erichson, *Arrhynchus* Philippi, *Exetasis* Walker and *Ocnaea* Erichson.

3.4. Timing of Acroceridae Evolution

We used a reduced nucleotide dataset of 65 loci to estimate a chronogram for spider flies. Fossilized birth-death (FBD) process divergence dating (Heath et al., 2014) performed here shows that the origin of crown spider flies dates back to the Upper Jurassic, at approximately 160 MYA (186–156 Ma 95% highest probability density interval, HPD) (Fig. 4, Supplementary Table 2). This new estimate for

the age of spider flies is much younger than the 198 MYA estimate recovered in a previous study (Winterton et al., 2007). Our results indicate that the major lineages of Acroceridae were already present by the Upper Cretaceous, but the greatest amount of cladogenesis occurred during the Paleogene, with most genera present by the end of that period. A few genera, however, evolved later in the Miocene, approximately 20–10 Ma ago (Fig. 4). The 95% HPD values for each node are given in Supplementary Table 2, and the numbered nodes in the Acroceridae phylogeny are presented in Supplementary Fig. 9. All 12 spider fly fossils (Gillung and Winterton 2017) were included as terminals in the dating analyses (Supplementary Table 3). The chronogram was very well supported overall, with all nodes (except for one) in the backbone of the tree with posterior probabilities (PP) of 1 (Fig. 4, Supplementary Table 2).

4. Discussion

4.1. Conflict among data types

We found surprising conflict between phylogenetic signal in the nucleotide and amino acid datasets, which resulted in two well-supported alternative hypotheses for spider fly evolution (Fig. 1). The fact that protein-coding gene sequence data (nucleotides) and their protein translations (amino acids) support conflicting phylogenies is highly significant since both types of data should have evolved under the same species tree as they are extracted from the same observations. The critical difference between the two topologies concerns the monophyly of the traditionally well-established and widely accepted subfamily Panopinae (Schlinger 1981, 2003; Winterton et al., 2007). The subfamily is recovered as monophyletic in analyses using nucleotide data, and as polyphyletic using amino acids (Fig. 1).

4.2. Reliability of the two alternative hypotheses

The topology based on nucleotides was far more robust to perturbations of the dataset, with results consistent when taxa with low gene occupancy are removed, third codon positions are excluded, loci are sampled based on evolutionary rate, and multiple phylogeny estimation methods are used (BI, ML and MSC). The topology based on amino acids, on the other hand, changes substantially when the data is perturbed, with nodes of interest in the backbone varying considerably (Fig. 1, Supplementary Figs. 4, 8). Moreover, the MSC analysis of the amino acid dataset suggests extensive levels of conflict among loci, a phenomenon that is further demonstrated in the extreme differences in tree topology if the fastest one third or slowest one third of the loci are analyzed separately (Supplementary Fig. 8).

We performed four-cluster likelihood mapping analysis (FcLM) to further understand the nature of the conflict among loci in the nucleotide and amino acid datasets. Results showed that the nucleotide dataset had much stronger support for the preferred topology, while in the amino acid dataset there was roughly equal support for the three alternative topologies (Fig. 2). Even though one topology was preferred in the amino acid dataset, its weight was not much greater than the weight towards the other two topologies (Fig. 2). When nucleotide loci were analyzed individually, a clear majority of genes supported the same topology that was preferred in the concatenated analysis (Fig. 2, Supplementary Table 1). In the case of the nucleotide dataset, when topology preference for each locus was weighted over its relative strength of support (see Supplementary Table 1), the support for the preferred tree was greater than the second preferred tree, while the third topology was supported by only a handful of genes. In contrast, when amino acid loci were analyzed individually, support was roughly equally split over the three possible topologies (Fig. 2, Supplementary Table 1). In summary, the nucleotide data was less equivocal regarding the preferred topology in both concatenated and individual loci analyses, while there was more conflict as to which of the topologies were preferred using amino acid data (Fig. 2, Supplementary Table 1).

Results of SymTest indicate that the amino acid and nucleotide datasets including 3rd codon positions violated, at least to some degree, the assumption of global stationarity, reversibility and homogeneity (SRH conditions) (Fig. 3). Violation of SRH conditions was much greater in the nucleotide dataset including 3rd codon positions, but when 3rd codon positions were removed, fewer violations were observed in the nucleotide dataset than in the amino acid dataset (Fig. 3). Since we obtained virtually the same tree topology when analyzing the nucleotide dataset with or without 3rd codon positions, this indicates that violation of SRH conditions was unlikely to strongly impact on our results. Thus, the conflict between the amino acid and the nucleotide topologies is likely not linked to SRH violation as measured by SymTest. It is generally assumed that phylogeny estimation based on nucleotides generally performs worse than amino acids specifically because nucleotides tend to violate SRH conditions to a greater extent than amino acids (Zwick et al., 2012; Rota-Stabelli et al., 2013; Cox et al., 2014). Nonetheless, we found evidence supporting the opposite case in our study. The removal of 3rd positions from the nucleotide dataset heavily reduced violation of SRH conditions. Thus, if tree estimation based on nucleotides was affected by violation of SRH conditions while that based on amino acids was not, the expectation is that after the removal of 3rd positions the resulting topology should be congruent with the amino acid tree.

When synonymous changes in the nucleotide dataset were masked using the *degen1* approach, the topologies based on the original and degenerated datasets were critically discordant (Fig. 1, Supplementary Figs. 4–6). These results suggest that there may be conflict in phylogenetic signal

originating from synonymous and nonsynonymous changes. We compared the overall number of variable sites in the original and degenerated nucleotide datasets and observed that the overall proportion of variable sites in 1st and 3rd codon positions decreased considerably when synonymous changes were masked. Overall proportion of variable sites decreased from 50% in the complete nucleotide alignment to only 14% in the degenerated dataset (Supplementary Fig. 10). Synonymous substitutions in 1st codon positions may be contributing the phylogenetic signal supporting the monophyly of Panopinae, because when synonymous changes are excluded, the resulting topology supports a non-monophyletic Panopinae (Supplementary Figs. 5–6).

Determining whether analysis of synonymous versus non-synonymous changes is likely to be more inaccurate is not trivial, especially in light of the absence of gross SRH violations in both cases, as measured by SymTest. We speculate that phenomena, including different selection regimes and different patterns of non-independence among sites, may result in nucleotide and amino acid sequences that subtly violate the assumptions of common phylogenetic models, which could affect inference based on synonymous and non-synonymous changes in different directions. Additionally, our FcLM results suggest differences in patterns of topological conflict among loci in the nucleotide and amino acid datasets. Gillung and Khouri et al. (in prep) are using posterior predictive simulation (Bollback 2002; Doyle et al., 2015; Duchene et al., 2016) to evaluate absolute model fit to the current datasets and investigating whether model misspecification is the source of conflict among and within the datasets.

The conflict within the results based on amino acids may have biological or methodological causes, including, for instance, incomplete lineage sorting and poor model fit to some genes or subsets of data, respectively. In either case, this decreases the credibility of the topology inferred from the amino acid concatenated dataset. If conflict among gene trees is real and pervasive, not modelling it explicitly could result in inaccurate estimates of topology. Species tree estimation methods can account for some of the biological sources of conflict; however, despite favoring a topology similar to that inferred from the concatenated dataset, our ASTRAL results are inconclusive due to the low support for the nodes of interest.

Results based on analyses of nucleotides were more robust to alterations of data and different analytical methods. This implies that either the results are accurate, or that there is pervasive systematic error affecting all nucleotide analyses (and most genes or data subsets) in the same way. Violation of SRH conditions is thought to be the most common source of error disproportionately affecting nucleotide analyses. Given that our SymTest results suggest that this is not the case for our dataset, we prefer the hypothesis of a monophyletic Panopinae, until further investigation.

4.3. Patterns of Acroceridae evolution

This study comprises the first phylogenomic treatment of Acroceridae relationships, with molecular sequence data sampled across all major spider fly lineages. Analyses of nucleotide data converged upon a fully resolved and well-supported tree topology that is incongruent with traditional hypotheses based on morphology (Schlinger 1987) and smaller sampling of molecular data (Winterton et al., 2007). Unprecedented aspects of our results include the placement of the morphologically derived, species-rich genus *Ogcodes* as sister to the rest of the Acroceridae other than *Acrocera* + *Carvalhoa*, and the non-monophyly of some traditionally well-established genera, including *Parahelle* and *Ocnaea* (Fig. 1). More importantly, the non-monophyly of Acrocerinae, already suggested by Winterton et al. (2007), indicates pervasive and strong discordance between traditional morphological systematics and molecular phylogenetic results.

Acrocerinae were polyphyletic in all of our analyses, with four independent clades. A clade composed of the enigmatic Chilean genus *Carvalhoa* and the cosmopolitan, species-rich *Acrocera* was recovered as sister to all other spider flies, in general agreement with previous molecular results (Winterton et al., 2007). Whilst adult *Acrocera* and *Carvalhoa* are morphologically similar to other Acrocerinae, having relatively small heads, bulbous bodies and reduced wing venation, their larval morphology and behaviour contrast with the rest of the family. Species in these two genera have unique associations with araneomorph spiders in the Haplogynae, while the remaining acrocerids attack Entelegynae araneomorph spiders or Mygalomorphae spiders (Gillung and Borkent 2017). Additionally, the first instar planidial larvae of all other acrocerids have well sclerotized body segments with setae or scales allowing them to actively locomote via looping, leaping and flicking movements (King 1916; Schlinger 1960b, 2003). Instead, *Acrocera* first instar larvae lack both sclerotization and long setae, and only crawl (Overgaard Nielsen et al., 1999).

The placement of *Ogcodes* as sister to the remaining Acroceridae (excluding *Carvalhoa* + *Acrocera*) was recovered with strong support in all analyses irrespective of data type (nucleotides or amino acids), phylogeny inference method (BI, ML or MSC) and alternative taxon and gene sampling. The genus was previously placed in the former Acrocerinae based on morphology (Schlinger 1987) and Sanger sequence data, albeit with low confidence (Winterton et al., 2007). This placement is justifiable as species of *Ogcodes* have small body size and reduced wing venation, highly apomorphic traits that are likely interrelated.

The two remaining clades of the former Acrocerinae comprise the genera *Turbopsebius* + *Cyrtus* and *Pterodontia* + *Psilodera*. Schlinger (1972) postulated the *Cyrtus–Opsebius* (including *Turbopsebius*) lineage of Acroceridae, and our results confirm their close relationship. The phylogenetic position of both *Pterodontia* and *Psilodera* has always been contentious (Schlinger 1960a, 1972; Winterton et al., 2007).

Psilodera was previously affiliated with the acrocerines *Pterodontia* and *Ogcodes*, although with weak statistical support (Winterton et al., 2007). Here, *Psilodera* and *Pterodontia* were recovered as the sister clade to the Panopinae in the nucleotide topology.

Monophyly of the bizarre Philopotinae has never been contested, and the clade was, unsurprisingly, recovered with strong statistical support in all of our analyses. Several morphological features define the subfamily, including enlarged postpronotal lobes forming a collar around the head and a distinct arched body shape (Schlinger 1987). Our phylogenomic analyses also strongly support the internal arrangement of Philopotinae as proposed by Winterton et al. (2007), with two main clades recovered. The first clade includes the genera *Megalybus*, *Philopota* and *Oligoneura*, with *Parahelle* and *Thyllis* in the second clade.

In all nucleotide-based topologies, Panopinae are monophyletic with high statistical support. Overall, reciprocal monophyly of individual panopine genera is well supported, except for *Exetasis* and *Ocnaea*. Schlinger (1968) differentiated the two genera based on two weak morphological characters, distribution of the microtrichia on the wing membrane and the absence of the wing vein R_4 in *Exetasis*, though other authors have dissented. Our results indicate that the two genera should probably be synonymized.

4.4. Timeline of Acroceridae evolution and diversification

The origin of Acroceridae has been estimated by Wiegmann et al. (2003) at approximately 175–225 MYA, and by Winterton et al. (2007) at *ca*. 173–221 MYA. Our results indicate a much younger age for the origin of the family in the Middle to Late Jurassic (156–187 MYA). This difference in age estimates might be due to a different calibration approach used here (tip dating versus node dating), greater fossil sampling, and a revised, younger age estimate of Baltic amber (Aleksandrova and Zaporozhets 2008). The age and plesiomorphic appearance of the oldest definitive spider flies, *Archocyrtus gibbosus* Ussatchov and *A. kovalevi* (Nartshuk), both described from late Jurassic fossil beds from Karatau, Kazakhstan (Ussatchov 1968; Nartshuk 1996), are consistent with a late Mesozoic origin and Cretaceous diversification of Acroceridae.

Whilst *Carvalhoa+Acrocera* diverged from the rest of the family relatively early (156–174 MYA, Middle Jurassic), the rest of Acroceridae radiated more recently, with the divergence of *Ogcodes* from the rest of Acroceridae occurring 45 million years later, in the Lower Cretaceous. Philopotinae diverged from Panopinae and remaining Acrocerinae in the Lower Cretaceous (97–131 MYA). Within Philopotinae, the New World and Oriental genera (*Megalybus*, *Philopota* and *Oligoneura*) diverged from Afrotropical genera (*Parahelle* and *Thyllis*) approximately 72–103 MYA, towards the Upper Cretaceous. Finally,

Acrocerinae *partim* and Panopinae diverged during the Middle Cretaceous (82–114 MYA), with crown group Panopinae appearing during the Upper Cretaceous (*ca*. 98–69 MYA) (Fig. 4, Supplementary Table 2).

5. Conclusion

We applied a phylogenomic approach to resolve the phylogeny of spider flies, sampling molecular sequence data of 240 homologous genes from all major lineages. Analyses of supermatrices as well as species tree approaches converged upon a robust hypothesis of Acroceridae evolution based on nucleotides under a variety of analytical parameters.

Acroceridae is remarkable within Diptera as the only parasitoid group specialized in spiders, with remaining fly parasitoids mainly attacking other insects. We took advantage of the recent advances in divergence time estimation (Heath et al., 2014) to propose a robust hypothesis for the timing of spider fly evolution, in which all known fossil acrocerids were included as terminals, with ages ranging from the Upper Jurassic to the Miocene (Gillung and Winterton 2017). The lack of clarity concerning the position of Acroceridae within the Diptera tree of life, however, limits how we can interpret the timing of the diversification of this specialized group of spider endoparasitoids.

Although more sequence data have often been shown to help resolve difficult phylogenetic questions, our study of spider fly phylogeny shows that simply increasing the amount of data can in fact be detrimental if added sequences have properties that introduce conflict among data types. When large-scale data matrices are used to study challenging nodes in the tree of life, relatively subtle model violations may be sufficiently amplified to mislead analyses, and those violations may not be obvious in many datasets. Thus, the comparative and exploratory approach implemented here may be a desirable way to detect conflicting signal in phylogenomic analyses. Specifically, it is important to compare topologies based on both amino acids and nucleotides because, even though they represent merely alternative coding of the same underlying data, their statistical analyses are fundamentally different (Huelsenbeck et al., 2008). In particular, the customary use of empirical amino acid models in which all of the potentially free parameters are fixed to specific values may be a source of model violation. Also, the use of global exchangeability rates as implemented in the CAT+GTR+G model might introduce tremendous amounts of model misspecification, because under this model it is assumed that all partitions share the same exchangeability rates.

Our results further provide an insight into the question of data type choice in phylogenomics and the importance of analyzing data both as amino acids and nucleotides. Careful analyses of data are critical, especially when larger amounts of sequence data are becoming available for inclusion in

phylogenetic studies. Exploratory analyses such as tests of compositional heterogeneity, posterior predictive approaches to assess absolute model fit (Bollback 2002; Doyle et al., 2015; Duchene et al., 2016), or sensitivity of results to removal of sites likely to introduce systematic error (Salichos and Rokas 2013; Goremykin et al., 2015) should become a part of the standard phylogenomics toolkit. In addition, future work on phylogenomics should focus on better understanding of how different biases affect different data sources.

Acknowledgements

This work was supported by the U.S. National Science Foundation (DEB-1144119 to SLW), and by the Brazilian National Council for Scientific and Technological Development (grant 209447/2013-3 to JPG). Thank you to the following people from the 1KITE team: Lars Podsiadlowski, Alexander Donath, Daniela Bartel, Sabrina Simon, Karen Meusemann, and to the 1KITE Antliophora group for providing transcriptome data (http://www.1kite.org/subprojects.html). Thank you to Lars Jermiin for making SymTest available for the authors. Thank you to Silvio S. Nihei and Carlos E. Lamas for the loan of material from the Museum of Sao Paulo (MZSP) for DNA sequencing. Also, thank you to Michelle Trautwein for contributing analytical organization and capability for the Wiegmann Lab. Thank you to Brian Cassel for his help with targeted enrichment and initial processing of sequences. Comments from Brendon B. Boudinot, Phillip S. Ward, Brian Moore, Karen Meusemann and Alfried Vogler refined some of the ideas presented here.

Supplementary Material

Supplementary Figures and Tables accompany this paper at the Zenodo Database (https://doi.org/10.5281/zenodo.1289998).

References

Ababneh, F., Jermiin, L.S., Ma, C., Robinson, J., 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22, 1225–1231.

Aberer, A.J., Kobert, K., Stamatakis, A., 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31, 2553–2556.

Adachi, J., Hasegawa, M., 1996. MOLPHY v.2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28, 1–150.

Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358.

Aleksandrova, G.N., Zaporozhets, N.I., 2008. Palynological characteristics of Upper Cretaceous and Paleogene deposits on the west of the Sambian Peninsula (Kaliningrad region), part 1. Stratigr. Geol. Correl. 16, 295–316.

Andrews, S., 2010. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bininda-Emonds, O.R.P., 2007. Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evol. Bioinform. 3, 59–85.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bollback, J.P., 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19, 1171–1180.

Borkent, C.J., Gillung, J.P., Winterton, S.L., 2016. Jewelled spider flies of North America: a revision and phylogeny of *Eulonchus* Gerstaecker (Diptera, Acroceridae). ZooKeys 619, 103–146.

Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660.

Borowiec, M.L., 2017. Convergent evolution of the army ant syndrome and congruence in big-data phylogenetics. BioRXiv, https://doi.org/10.1101/134064.

Borowiec, M.L., Lee, E.K., Chiu, J.C., Plachetzki, D.C., 2015. Dissecting phylogenetic signal and accounting for bias in whole-genome data sets: a case study of the Metazoa. Mol. Biol. Evol. 16, 987.

Bowker, A.H., 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43, 572–574.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10, e1003537.

Branstetter, M.G., Danforth, B.N., Pitts, J.P., Faircloth, B.C., Ward, P.S., Buffington, M.L., Gates, M.W., Kula, R.R., Brady, S.G., 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27, 1019–1025.

Cady, A., Leech, R., Sorkin, L., Stratton, G., Caldwell, M., 1993. Acrocerid (Insecta: Diptera) life histories, behaviors, host spiders (Arachnida: Araneida), and distributional records. Can. Entomol. 125, 931–944.

Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S., Hasegawa, M., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.

Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA 112, 14912–14917.

Cox, C.J., Li, B., Foster, P.G., Embley, T.M., Civán, P., 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol. 63, 272–279.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8, 783–786.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins, in: Dayhoff, M.O. (Ed.), Atlas of protein sequence and structure, Vol. 5, Supplement 3. National Biomedical Research Foundation, Washington DC, pp. 345–352.

Doyle, V.P., Young, R.E., Naylor, G.J.P., Brown, J.M., 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64, 824–837.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.

Duchene, S., Di Giallonardo, F., Holmes, E.C., 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales Mol. Biol. Evol. 33, 255–267.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Espeland, A., Breinholt, M., Willmott, J., Warren, K.R., Vila, A.D., Toussaint, R., Maunsell, E.F.A., Aduse-Poku, S.C., Talavera, K., Eastwood, G., et al., 2018. A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Curr. Biol. 28, 770–778.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Fučíková, K., Lewis, P.O., Lewis, L.A., 2016. Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution. Mol. Phylogenet. Evol. 98, 176–83.

Garrison, N.L., Rodriguez, J., Agnarsson, I., Coddington, J.A., Griswold, C.E., Hamilton, C.A., Hedin, M., Kocot, K.M., Ledford, J.M., Bond, J.E., 2016. Spider phylogenomics: untangling the spider tree of life. PeerJ 4, e1719.

Gillung, J.P., Winterton, S.L., 2011. New genera of philopotine spider flies (Diptera, Acroceridae) with a key to living and fossil genera. ZooKeys 127, 15–27.

Gillung, J.P., Nihei, S.S., 2016. Evolution of Philopotinae, with a revision and phylogeny of the New World spider fly genus *Philopota* Wiedemann (Diptera, Acroceridae). Zool. J. Linnean. Soc. 176, 707–780.

Gillung, J.P., Borkent, C.J., 2017. Death comes on two wings: a review of dipteran natural enemies of arachnids. J. Arachn. 45, 1–19.

Gillung, J.P., Winterton, S.L., 2017. A review of fossil spider flies (Diptera: Acroceridae) with descriptions of new genera and species from Baltic Amber. J. Syst. Palaeontol. 16, 325–350.

Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, M., Lockhart, P., 2015. The root of flowering plants and total evidence. Syst. Biol. 64, 879–891.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Haddad, S., Shin, S., Lemmon, A.R., Lemmon, E.M., Svacha, P., Farrell, B.D., Slipinski, A., Windsor, D., McKenna, D.D., 2018. Anchored hybrid enrichment provides new insights into the phylogeny and evolution of longhorned beetles (Cerambycidae). Syst. Ent. 43, 68–89.

Hamilton, C.A., Lemmon, A. ., Lemmon, .M., Bond, . ., 2016. panding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol. 16, 212.

Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. USA 111, E2957–E2966.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.

Huelsenbeck, J.P., Joyce, P., Lakner, C., Ronquist, F., 2008. Bayesian analysis of amino acid substitution models. Philos. Trans. R. Soc. B 363, 3941–3953.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.

Jermiin, L., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53, 638–643.

Jermiin, L.S., Jayaswal, V., Ababneh, F., Robinson, J., 2008. Phylogenetic model evaluation, in: Keith, J.M. (Ed.), Bioinformatics, Volume 1: Data, Sequence Analysis, and Evolution. Humana Press, Totowa, pp. 331–364.

Jia, F., Lo, N., Ho, S.Y.W., 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS ONE 9, e95722.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods 14, 587–589.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

King, J.L., 1916. Observations on the life history of Pterodontia flavipes Gray (Diptera). Ann. Entomol. Soc. Am. 9, 309–321.

Kocot, K.M., Struck, T.H., Merkel, J., Waits, D.S., Todt, C., Brannock, P.M., Weese, D.A., Cannon, J.T., Moroz, L.L., Lieb, B., et al. 2016. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst. Biol. 66, 256–282.

Kozlov, A.M., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31, 2577–2579.

K c k, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W., Misof, B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7, 10.

Kutty, S.N., Wong, W.H., Meusemann, K., Meier, R., Cranston, P.S.. 2018. A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. Syst. Ent. 35, 823–836.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14, 82.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.

Lartillot, N., Philippe, H., 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Syst. Biol. 21, 1095–1109.

Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7, S4.

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

Le, S.Q., Dang, C.C., Gascuel, O., 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29, 2921–2936.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Mayrose, I., Friedman, N., Pupko, T.. 2005. A gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21, 151–158.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop, New Orleans.

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes Bioinformatics 31, i44-i52.

Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T., 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346, 250463–1250463.

Misof, B., Misof, K.A., 2009. Monte Carlo approach successfully identifies randomness of multiple sequence alignments: a more objective approach of data exclusion. Syst. Biol. 58, 21–34.

Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K., Meusemann, K., 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics 14, 348.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767.

Nabholz, B., Künstner, A., Wang, R., Jarvis, E.D., Ellegren, H., 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28, 2197–2210.

Nartshuk, E.P., 1996. A new fossil acrocerid fly from the Jurassic beds of Kazakhstan (Diptera: Acroceridae). Zoosystematica Rossica 4, 313–315.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Overgaard Nielsen, B., Funch, P., Toft, S., 1999. Self-injection of a dipteran parasitoid into a spider. Naturwissenschaften 86, 530–532.

Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., et al.. 2017. Evolutionary history of the Hymenoptera. Curr. Biol. 27, 1013–1018.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., et al., 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics 18, 111.

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA 112, 15402–15407.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A.J., 2014. Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer/.

Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.L., Harshman, J., Huddleston, C.J., Kingston, S., et al., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66, 857–879.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.

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

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D., Rasnitsyn, A.P., 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61, 973–999.

Rota-Stabelli, O., Lartillot, N., Philippe, H., Pisani, D., 2013. Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. 62, 121–133.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol. Biol. Evol. 30, 197–214.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signal. Nature 497, 327–331.

Schlinger, E.I., 1960a. A review of the South African Acroceridae (Diptera). Ann. Natal Museum 14, 459–504.

Schlinger, E.I., 1960b. A review of the genus *Eulonchus* Gerstaecker. Part I. The species of the smaragdinus group (Diptera: Acroceridae). Ann. Am. Entomol. Soc. 53, 416–422.

Schlinger, E.I., 1968. A revision of *Arrynchus* Philippi and a key to the genera of the *Ocnaea* branch of the Panopinae (Diptera). Rev. Chil. Entomol. 6, 47–54.

Schlinger, . I., 1972. New east Asian and American genera of the "*Cyrtus*-*Opsebius*" branch of the Acroceridae (Diptera). Pacific Insects 14, 409–428.

Schlinger, E.I., 1981. Acroceridae, in: McAlpine, J.F., Peterson, B.V., Shewell, G.E., Teskey, H.J., Vockeroth, J.R., Wood, D.M. (Eds.), Manual of Nearctic Diptera. Vol. 1. Agriculture Canada Research Branch, Monograph 27, Ottawa, pp. 575–584.

Schlinger, E.I., 1987. The biology of Acroceridae (Diptera): true endoparasitoids of spiders, in: Nentwig, W. (Ed.), Ecophysiology of Spiders. Springer-Verlag, Germany, pp. 319–327.

Schlinger, E.I., 2003. Acroceridae, spider endoparasitoids, in: Goodman, S.M., Benstead, J.P. (eds.), The Natural History of Madagascar. University of Chicago Press, Chicago, pp. 734–740.

Schlinger, E.I., Gillung, J.P., Borkent, C.J., 2013. New spider flies from the Neotropical Region (Diptera, Acroceridae) with a key to New World genera. Zookeys 270, 59–93.

Shaffer, H., Minx, P., Warren, D., Shedlock, A.M., Thomson, R.C., Valenzuela, N., Abramyan, J., Badenhorst, D., Biggar, K.K., Borchert, G.M., et al., 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14, R28.

Shin, S., Clarke, D.J., Lemmon, A.R., Lemmon, E.M., Aitken, A.L., Haddad, S., Farrell, B.D., Marvaldi, A.E., Oberprieler, R.G., McKenna, D.D., 2017. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. Mol. Biol. Evol. 35, 823–836.

Shin, S., Bayless, K.M., Winterton, S.L., Dikow, T., Lessard, B.D., Yeates, D.K., Wiegmann, B.M., Trautwein, M.D., 2018. Taxon sampling to address an ancient rapid radiation: a supermatrix phylogeny of early brachyceran flies (Diptera). Syst. Ent. 43, 277–289.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Streicher, J.W., Schulte, J.A., Wiens, J.J., 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65, 128–145.

Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci USA 94, 6815–6819.

Sullivan, J., Swofford, D.L., Naylor, G.J.P., 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16, 1347–1356.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, 609–612.

Ussatchov, D.A., 1968. New Jurassic Asilomorpha (Diptera) fauna from Karatau. Entomol. Rev. 47, 617–628.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666.

Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., Li, C., White, S., Xiong, Z., Fang, D., et al., 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. 45, 701–706.

Waterhouse, R.M., Tegenfeldt, F., Zdobnov, E.M., Kriventseva, E.V., 2013. OrthoDB: a hierarchical catalog of animal fungal and bacterial orthologs. Nucleic Acids Res. 41, 358–365.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Wiegmann, B.M., Thorne, J.L., Yeates, D.K., Kishino, H., 2003. Time flies: A new molecular time-scale for fly evolution without a clock. Syst. Biol. 52, 745–756.

Wiegmann, B.M., Trautwein, M.D., Winkler, I.S., Barr, N.B., Kim, J.W., Lambkin, C., Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., et al., 2011. Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. USA 108, 5690–5695.

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. Mol. Phylogenet. Evol. 43, 808–832.

Winterton, S.L., Gillung, J.P., 2012. A new species of spider fly in the genus *Sabroskya* Schlinger from Malawi, with a key to Acrocerinae world genera (Diptera, Acroceridae). Zookeys 171, 1–15.

Winterton, S.L., Lemmon, A.R., Gillung, J.P., Garzon, I.J., Badano, D., Bakkes, D.K., Breitkreuz, L.C.V., Engel, M.S., Lemmon, E.M., Liu, X., et al., 2018. Evolution of lacewings and allied orders using anchored phylogenomics (Neuroptera, Megaloptera, Raphidioptera). Syst. Entomol. 43, 330–354.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evolut. 11, 367–372.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.

Yeates, D.K., Meusemann, K., Trautwein, M., Wiegmann, B., Zwick, A., 2016. Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution. Curr. Opin. Insect Sci. 13, 16–23.

Young, A.D., Lemmon, A.R., Skevington, J.H., Mengual, X., Ståhls, G., Reemer, M., Jordaens, K., Kelso, S., Lemmon, E.M., Hauser, M., et al., 2016. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). BMC Evol. Biol. 16, 143.

Zwick, A., Regier, J.C., Zwickl, D.J., 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. PLoS ONE 7, e47450.

Figure Legends

Figure 1. Phylogeny of spider flies based on the nucleotide (A) and amino acid (B) alignments. Green circles indicate nodes with posterior probability lower than 0.99 and/or bootstrap values lower than 80.

Figure 2. Four-cluster likelihood mapping (FcLM) analyses results. **A**. Phylogram of Acroceridae based on nucleotides showing the four taxon clusters used. **B**. FcLM results for each individual locus in the nucleotide and amino acid datasets. Bars show the percentage of loci supporting each of the three possible unrooted topologies, with darker colors (bars on left) showing raw percentages, and lighter colors (bars on left) showing percentages weighted over relative support for each topology as shown in Supplementary Table 1. **C**. FcLM results for the concatenated nucleotide and amino acid alignments. Values at the corners indicate the percentage of fully resolved phylogenies for all possible quartets.

Figure 3. Heat maps showing the results from pairwise comparison of aligned amino acid and nucleotide sequences (with and without 3rd codon positions) using Bowker's matched-pairs tests of symmetry. Cells

in white specify p-values > 0.05, indicating that the corresponding pair of sequences seemingly does not violate the assumption of global stationarity, reversibility and homogeneity (SRH conditions).

Figure 4. Estimated divergence times among lineages of Acroceridae under the fossilized birth-death process, in BEAST 2. Scale is in MYA. Bars depict the 95% highest posterior probability density of each estimate. Mean ages and ranges are provided in Supplementary Table 3 and refer to nodes indicated in Supplementary Fig. 10.

Highlights

- We present the first phylogenomic scale study of a dipteran parasitoid family.
- Analyses of nucleotides and amino acids resulted in conflicting estimates of phylogeny.
- Common sources of bias often attributed to nucleotides were not identified here.
- The origin of spider flies is younger than previously proposed.
- We demonstrate the importance of analyzing data both as amino acids and nucleotides.