ABSTRACT

TURCATEL, MAUREN. A Molecular Phylogeny of Deer Flies and their Closest Relatives. (Under the direction of Dr. Brian M. Wiegmann).

Horse flies (Diptera: Tabanidae) are an economically, medically, and ecologically important group of insects. Currently, the most accepted classification of the Tabanidae divides the family into three subfamilies: Tabaninae, Pangoniinae and Chrysopsinae. The Chrysopsinae is the least species rich, with 577 species in 34 genera and 3 tribes worldwide: Bouvieromyiini, with 173 species in 12 genera, Chrysopsini, with 336 species in 9 genera, including the true deer flies *Chrysops* Meigen, and Rhinomyzini, with 68 species in 13 genera. Previous hypotheses based on morphological data showed a tendency for Bouvieromyiini and Chrysopsini to merge, and the Rhinomyzini to be polyphyletic. However, modern analyses strongly suggest that the subfamily Chrysopsinae is not a monophyletic group, and may be broken into at least four subfamilies. In the first chapter of this thesis, phylogenetic analysis based on standard mitochondrial and nuclear genes of deer flies is performed, aiming to test the monophyly and relationships of existing tribes in order to support a new classification at the subfamily level. A divergence times analysis is used to estimate the ages of origin for major lineages within the radiation of deer flies.

In the second chapter, maximum likelihood inference of the Chrysopsinae was performed using molecular data generated by a custom set of anchored enrichment probes, intending to achieve a better resolution of the phylogenetic relationships of deer files based on new methods for obtaining genes and the use of next-generation sequencing technology.

And finally, the third chapter consists of the application of transcriptome data from 10 species of brachyceran flies with relatively certain phylogenetic relationships to test a novel automated workflow (Agalma) as a tool for generating alignments of homologous genes and building species trees.

A Molecular Phylogeny of Deer Flies and their Closest Relatives

by
Mauren Turcatel

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Entomology

Raleigh, North Carolina

2014

APPROVED BY:

Dr. Brian M. Wiegmann Committee Chair	Dr. D. Wes Watson
Dr. Marcé Lorenzen	Dr. (Jenny) Qiu-yun Xiang

BIOGRAPHY

Mauren Turcatel is originally from Brazil; her hometown is Cascavel, in the State of Paraná, more than an hour from the Iguazu Falls in the border of Brazil and Argentina. Her career path in science started while she was pursuing a degree in Biological Sciences at the Federal University of Paraná (UFPR) and a degree in Printmaking at the School of Music and Fine Arts of Paraná (EMBAP) at the same time, driven by her passion for scientific illustrations. A zoology professor noticed her drawings during a lab practice and asked if she was interested in an internship; Mauren agreed, and then he introduced her to several professors at the Zoology and Entomology departments at UFPR. That was how she met her first advisor, Dr. Claudio J. B. de Carvalho, who supervised her work during 3 years as an undergraduate, and 2 years as a Masters student.

Mauren's research in collaboration with Dr. Carvalho was focused on the taxonomy of Neotropical horse flies. In her undergraduate studies, she surveyed the horse fly species of the State of Paraná and used her art skills to develop pictorial identification keys. Her Master's thesis focused on the revision of an Amazonian genus of horse flies and included the descriptions of 2 new species. Some of Mauren's colleagues from UFPR and Dr. Carvalho have worked in collaboration with Dr. Wiegmann in the past, and in 2010 she joined the Wiegmann lab in the Department of Entomology , North Carolina State University (NCSU) as a PhD student in the NSF-PEET project on horse fly systematics and taxonomy.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1:

CHAPTER 3:

LIST OF FIGURES

INTRODUCTION

The family Tabanidae, with approximately 4400 species described (Pape and Thompson 2012), is the largest family of bloodsucking insects (Downes 1971). Nearly all females are blood-feeders as adults and can be responsible for the mechanical transmission of parasites of human and veterinary significance (Krinsky 1976), such as *Loa loa* filariasis, tularemia, anaplasmosis, trypanosomiasis, and anthrax (Pechuman and Teskey 1981). Their mouthparts are strong and blade-like, hence the bites are extremely painful and cause blood loss and stress to livestock, domestic animals, and humans (Krinsky 1976, Fairchild 1986). Most species feed on large mammals (e.g., cattle and horses), while a minority feed on reptiles or birds (Ferreira 2004). All males and some females rely on pollen and nectar as a carbohydrate source, and for this reason horse flies are also important pollinators in some ecosystems (Johnson & Morita 2006; Morita 2011). Adult horse flies are large, heavy-bodied flies (usually between 6mm and 30mm in length), with wide heads, brightly colored eyes, and relatively long antennae (Pechuman and Teskey 1981). Tabanidae larvae are venomous, generalist predators, and are typically found in freshwater and semi-aquatic environments (Teskey 1990).

Because of their generalized morphology, and the scarcity of known male specimens, taxonomy of horse flies has been historically difficult (Oldroyd 1957, Usher 1972, Chainey 1993). The lack of definitive morphological characteristics for identifying genera and species makes horse flies difficult to classify, for instance, large structural differences in the genitalia do not correspond with differences in external morphology, thus complicating species identification (Mackerras 1955). Common diagnostic characters can also be compromised by preservation methods (e.g. color fading in ethanol preserved and pinned specimens), affecting species-level identification (Mackerras, 1960; Mackerras et al., 2008). Tabanidae is currently subdivided into three subfamilies, with the number of tribes varying between 9 and 11 (Mackerras 1954).

Pangoniinae is composed of robust flies, distributed worldwide, mostly known for their long proboscis and their important role as pollinators. Previous studies recovered Pangoniinae as a monophyletic group, strongly supported as the sister to all other horse flies (Morita et al., in press*;* Bayless and Wiegmann 2012*,* Lessard et al., 2013).

Tabaninae constitutes the majority of horse fly species diversity, including 1350 described species in the cosmopolitan *Tabanus* Linnaeus, and 555 species in *Haematopota* Meigen (Bayless and Wiegmann 2012). Tabaninae is distinguished from deer flies by the loss of the tibial spurs on the metathoracic leg (Mackerras. 1954), and is strongly supported as monophyletic (Wiegmann et al. 2000; Morita et al. submitted, Bayless and Wiegmann 2012).

And finally, the subfamily Chrysopsinae, commonly known as deer flies, are aggressive biters and often display intricate eye patterns and elongate antennae. The most species rich and widely distributed genera in this subfamily are *Chrysops* and *Silvius*, while the majority of genera are small, several monotypic, and limited in distribution (Mackerras 1955). The majority of studies regarding the Chrysopsinae are descriptive, with revisions and redescriptions for only a few specific areas (e.g. Philip and Mackerras 1959, Burger and Chainey 2000) and notes about geographic distribution (e.g. Burger 1974, Fairchild 1978). The taxonomy of this subfamily is challenging, and since there are no clear synapomorphies unifying the tribes, groups are often defined by plesiomorphic characters or, admittedly, by convenience (Chainey 1987). The most commonly accepted classification scheme remains the one proposed by Mackerras (1954), who, based on the shared possession of a divided ninth tergite, inferred that Chrysopsinae and Tabaninae were sister taxa. Recent molecular analyses, which included a few Chrysopsinae specimens, placed Chrysopsinae as intermediate between Pangoniinae and Tabaninae, suggesting it is not monophyletic (Morita 2008, Bayless and Wiegmann 2012).

Morphological (Mackerras 1954; Stuckenberg 2001) and molecular (Wiegmann et al. 2000) studies strongly support the monophyly of Tabanidae. Phylogenies below the family level were reconstructed in molecular studies focusing on the subfamilies Pangoniinae (Morita 2008, Morita et al. in press, Lessard et al. 2013) and Tabaninae (Bayless and Wiegmann 2012). As a part of a National Science Foundation Partnerships for Enhancing Expertise in Taxonomy (PEET) grant aimed at expanding the knowledge of horse flies, this project minutely investigates the phylogenetic relationships of the tribes and genera of Chrysopsinae. Here we use molecular characters of these species for the first time, seeking evidence for monophyletic lineages and aiming to provide a phylogenetic hypothesis as context for revisionary taxonomic studies.

In order to assess the evolutionary history of this group, this study includes a total of 10 genera of the three tribes of Chrysopsinae (*Silvius* and *Chrysops* Meigen, Chrysopsini; *Thaumastocera* and *Orgizomyia* Grunberg, and *Tabanocella* Bigot, Rhinomyzini; *Aegophagamyia* Austen, *Rhigioglossa* Wiedemann, *Lilaea* Walker, *Pseudotabanus* Ricardo, and *Coracella* Philip, Bouvieromyiini), along with 11 genera of the subfamily Pangoniinae (designated as outgroup) and 14 genera of Tabaninae previously sampled for other integrant projects of the PEET grant. Nine molecular markers were sampled and amplified using standard DNA extraction and PCR protocols, resulting in a concatenated dataset of 4586 nucleotides. Maximum Likelihood and Bayesian inference were performed, as well as a divergence times analysis to estimate the age of origin for major lineages within the radiation of deer flies.

In the second chapter, Anchored Enrichment (Lemmon et al. 2012) was used aiming to achieve a better resolution of the phylogenetic relationships of deer flies. In this approach, specific genomic regions are captured prior to sequencing; capture probes for the target regions are designed from known sequences, within highly conserved anchor regions with rapidly evolving adjacent regions to provide sufficient phylogenetic information (Lemmon et al. 2012). The probes are then mixed with genomic DNA either on an array or in solution, and only DNA that successfully hybridizes to these probes is sequenced using high-throughput technology (McCormack et al. 2013, Lemmon and Lemmon 2013, Hedtke et al. 2013). This is a promising approach for both shallow and deep level phylogenetic studies of non-model groups, without the need of a reference genome.

The third chapter focuses in the use of transcriptome data to infer deep-level phylogenies of flies. Transcriptome data is a cost-effective alternative to whole genome sequencing (Lemmon and Lemmon 2013), and is obtained by an approach called RNA-Seq, where RNA is extracted from specific tissues of the target organism, and then reverse transcribed into cDNA; each molecule is sequenced with high-throughput technology, and the resulting sequence reads can either be aligned with the reference genome or transcripts or assembled using *de novo* strategies, and for this reason, RNA-Seq is attractive for non-model organisms whose genomes have not been sequenced (Lemmon and Lemmon 2013, Wang et al. 2009). Here, we used available transcriptome data from 10 species of Tabanomorpha with previously known phylogenetic relationships to test Agalma (Dunn et al. 2013), a novel automated tool developed to construct matrices for phylogenetic analyses from transcriptome data.

REFERENCES

Bayless, K.M. and Wiegmann, B.M. (2012). *Evolutionary Studies in Tabanomorpha (Diptera) with New Classifications of Xylophagidae and Tabaninae.* Unpublished manuscript. Department of Entomology, North Carolina State University, Raleigh, North Carolina.

Burger, J.F. (1974). Horse Flies of Arizona II. Notes on and keys to the adult Tabanidae of Arizona, Subfamilies Pangoniinae and Chrysopsinae (Diptera). Proceedings of the Entomological Society of Washington 76(9): 247–269.

Burger, J.F., and Chainey, J.E. (2000). Revision of the Australasian species of Chrysops (Diptera: Tabanidae). Invertebrate Taxonomy 14: 607–654.

Chainey, J.E. (1987). Afrotropical Tabanidae (Diptera): the genus Rhigioglossa Wiedemann, 1828 (including *Mesomyia* Macquart, 1850, as a subgenus). *Annals of the Natal Museum 28*: 137-159.

Chainey, J.E. (1993). Horse–flies, deer-flies and clegs (Tabanidae). In: Lane PR, Crosskey RW. (Eds) Medical Insects and Arachnids. Chapman & Hall, London, 706pp.

Downes, J.A. 1971. The ecology of blood-sucking Diptera: an evolutionary perspective. pp. 232-258 in Fallis, A.M. (Ed.) Ecology and physiology of parasites. Toronto, Ontario, University of Toronto Press.

Dunn, C.W., Howison, M., and Zapata, F. (2013). Agalma: an automated phylogenomics workflow. BMC *Bioinformatics 14*: 330.

Fairchild, G.B. (1978). New and little known Florida Tabanidae. The Florida Entomologist 61(3): 121–138.

Fairchild, G.B. (1986). Tabanidae of Panama. *Contributions of the American Entomological Institute 22*, 1–139.

Ferreira, R.L.M. and Rafael, J.A. (2004). Activities of tabanids (Diptera, Tabanidae) attacking domestic duck *Cairina moschata* (Linnaeus) (Aves, Anatidae), introduced in a forest area in the Central Amazon, Manaus, Brazil. *Revista Brasileira de Entomologia 48*(2): 283-286.

Hedtke, S.M., Morgan, M.J., Cannatella, D.C., & Hillis, D.M. (2013). Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time. *PLOS ONE 8*(7): e67908.

Johnson, S.D., Morita, S. (2006). Lying to Pinocchio: floral deception in an orchid pollinated by long-proboscid flies. *Botanical Journal of the Linnean Society 152*: 271–278.

Krinsky, W. L. (1976). Animal disease agents transmitted by horse flies and deer flies (Diptera: Tabanidae). *Journal of Medical Entomology 13*: 225–275

Lemmon, A.R., Emme, S.A., & Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology 61*: 727–744.

Lemmon, E.M. and Lemmon, A.R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Reviews of Ecology, Evolution and Systematics 44*: 99–121.

Lessard, B. D. and Yeates, D. K. (2012). Seven new Australian species of the Southern hemisphere horse fly subgenus *Scaptia (Pseudoscione)* (Diptera: Tabanidae), including descriptions and a revised key. *Journal of Medical Entomology 49*(6): 1206–1225.

McCormack, J.E., Hird, S.M., Zellmer A.J., Carstens, B.C., Brumfield, R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution 66*: 526–38.

Mackerras, I. M. (1954). The classification and distribution of Tabanidae (Diptera) I. General Review. *Australian Journal of Zoology 2*: 431–454.

Mackerras, I. M. (1955). The classification and distribution of Tabanidae (Diptera) III. Subfamilies Sepsidinae and Chrysopsinae. *Australian Journal of Zoology 3*: 583–633.

Mackerras, I.M., Spratt, D.M., Yeates, D.K. (2008). Revision of the horse fly genera Lissimas and Cydistomyia (Diptera: Tabanidae: Diachlorini) of Australia. Zootaxa 1886, 1–80.

Mackerras, I.M. (1960). The Tabanidae (Diptera) of Australia III. Subfamily Pangoniinae, tribe Scionini and supplement to Pangoniini. *Australian Journal of Zoology 8*: 1–152.

Morita, S.I. (2008). A phylogeny of long-tongued horse flies (*Philoliche*, Diptera: Tabanidae) with the first cladistic evaluation of higher relationships within the family. *Invertebrate Systematics 22*(3): 311–327.

Morita, S. I. (2011). Repeatability and precision in proboscis length measurements for long proboscid flies. *Zootaxa 3112*: 49-58.

Oldroyd, H. (1957). The Horse-flies (Diptera: Tabanidae) of the Ethiopian Region III: Chrysopsinae, Scepsidinae, and Pangoniinae and a Revised Classification. British Museum, London. 489 pp.

Pape, T., & Thompson, F. C. (2012). Systema dipterorum (version 2.0, Jan 2011). In: Bisby, F., Roskov, Y., Culham, A., Orrell, T., Nicolson, D., Paglinawan, L., Bailly, N., Kirk, P., Bourgoin, T., Baillargeon, G. (Eds.), Species 2000 & ITIS Catalogue of Life. Digital resource at www.catalogueoflife.org/col/. Species 2000, Reading, UK. Accessed (05.13.13)

Pechuman, L.L. & Teskey, H.J. (1981). Tabanidae. In: Manual of Neartic Diptera (J. F. McAlpine et al. ed.). Research Branch Agriculture Canada, Ottawa, v.1, p.464-478.

Philip, C.B., and Mackerras, I.M. (1959). On Asiatic and related Chrysopinae (Diptera: Tabanidae). The Philippine Journal of Science 88: 279–324.

Stuckenberg, B. R. (2001). Pruning the tree: a critical review of classifications of the Homeodactyla (Diptera, Brachycera), with new perspectives and an alternative classification. *Studia Dipterologica 8*(1), 3–41.

Teskey, H.J. (1990). The Insects and Arachnids of Canada. Part 16. The Horse Flies and Deer Flies of Canada and Alaska. Diptera: Tabanidae. Publication 1838. Research Branch, Agriculture Canada, Ottawa. 381pp.

Usher, P. J. (1972). A review of the South African horsefly fauna (Diptera: Tabanidae). *Annals of the Natal Museum 21*: 459–507

Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 10*: 57–63.

Wiegmann, B.M., Tsaur, S.C., Webb, D.W., Yeates, D. K., and Cassel, B. K. (2000). Monophyly and relationships of the Tabanomorpha (Diptera: Brachycera) based on 28S ribosomal gene sequences. *Annals of the Entomological Society of America 93*(5): 1031–1038.

CHAPTER 1: Phylogenetic relationships of flies in the tabanid subfamily Chrysopsinae

INTRODUCTION

The family Tabanidae comprises about 4400 extant described species throughout the world (Pape and Thompson 2012). Nearly all females are blood-feeders as adults and some can be responsible for the mechanical transmission of parasites of human and veterinary significance (Krinsky 1976). Most species feed on large mammals (e.g., cattle and horses), while a minority feed on reptiles or birds (Ferreira 2004). Interestingly, horse flies are also important pollinators, as all males and some females feed exclusively on nectar and pollen (Lessard and Yeates 2012).

Traditionally, the family Tabanidae is subdivided into three main subfamilies: Tabaninae, Pangoniinae and Chrysopsinae. Of these, the subfamily Chrysopsinae, containing the deer flies, is the least species rich, with 580 species in 35 genera, and 3 tribes worldwide: Bouvieromyiini, Chrysopsini, and Rhinomyzini (appendix A). Most Chrysopsinae species can be recognized by the presence of hind tibial spurs, well-developed ocelli, and antennal flagellum with 5 or fewer segments (Mackerras 1955).

The tribe Bouvieromyiini (147 spp., 13 genera) is a heterogeneous group of flies, most diverse in Madagascar, that vary in their genital morphology in such manner that sometimes the two halves of ninth tergite remain narrowly joined at the midline (Mackerras 1955). Afrotropical species previously described as *Silvius* (Chrysopsini) are now recognized as *Aegophagamyia*.

The tribe Chrysopsini (336 spp., 8 genera) includes the large, cosmopolitan genera *Silvius* and *Chrysops*, a well-known genus with intricately patterned eyes. Other genera in this group have relatively few species, several monotypic, and limited in distribution.

The tribe Rhinomyzini (69 spp., 14 genera) contains genera with highly specialized morphology such as enlarged palpi and complex antennal projections, and are further differentiated by the possession of a sub-apical ridge in the male gonostyle

(Mackerras 1954). This tribe is well represented in tropical Africa and Southeast Asia. There is only one representative of the tribe Rhinomyzini in the New World, *Betrequia ocellata* Oldroyd 1970, a species widely distributed in the Amazon Basin in Brazil. Hematophagy is not known to occur in most Rhinomyzini species from the Old World (Oldroyd 1957), and the Amazonian species was observed feeding only on caiman crocodiles among different hosts (caiman, horse, duck and human), suggesting that crocodilians may be its natural and ancestral host (Henriques et al. 2000).

The monophyly of the Tabanidae is well supported by both molecular (Wiegmann et al. 2000) and morphological evidence (Mackerras 1954; Stuckenberg 2001). Molecular analyses support the monophyly of the subfamilies Pangoniinae and Tabaninae, but these studies included only a few specimens of Chrysopsinae, which were positioned as intermediate between Pangoniinae and Tabaninae and not recovered as a monophyletic group, hence suggesting only two monophyletic subfamilies of Tabanidae or the delineation of four or more (Morita 2008, Bayless and Wiegmann 2012).

There are no clear morphological synapomorphies unifying the three tribes of Chrysopsinae, hence molecular data is of critical importance to assess the evolutionary history of this group. This study aims to test the monophyly and relationship of the three existing tribes of the paraphyletic subfamily Chrysopsinae based on molecular characters for the first time, in order to provide evidence for monophyletic lineages and a phylogenetic hypothesis as context for revisionary taxonomic studies. Divergence time analyses will be used to estimate ages of major lineages within the radiation of deer flies.

MATERIAL AND METHODS

Taxon sampling

Specimens preserved in ethanol from several collecting trips in different locations obtained by previous students of the Tabanidae PEET grant and other collaborators were available at the Wiegmann lab.

Ten genera representing the three tribes of Chrysopsinae were sampled: three species of *Silvius* and 27 *Chrysops* Meigen (Chrysopsini); one *Thaumastocera* and one *Orgizomyia* Grunberg, and three *Tabanocella* Bigot (Rhinomyzini); four *Aegophagamyia* Austen*,* three *Rhigioglossa* Wiedemann*,* two *Lilaea* Walker*,* two *Pseudotabanus* Ricardo*,* and one *Coracella* Philip*,* (Bouvieromyiini). Ideally a more thorough taxon sampling would have included all 35 Chrysopsinae genera; however, several genera are either monotypic or have very few described species with limited geographical distributions. Also, a significant fraction of our samples are from tropical areas, mainly Madagascar, and show signs of DNA degradation probably caused by excessive heat and humidity.

Fifteen species of 12 genera of Pangoniinae and 14 species representing 14 genera of Tabaninae were also included. The Pangoniinae were designated as the outgroup, since these flies are strongly supported as a monophyletic group and as the sister to all other horse flies (Morita 2008*,* Lessard et al., 2013).

DNA extraction, amplification, and sequencing

Genomic DNA was extracted using the DNeasy DNA extraction kit (QIAGEN, Inc.). The standard protocol was altered by extending the amount of time the specimen was in proteinase K solution to 2 days to allow more cells to be broken down. The final elution was reduced to 30 μL to avoid diluting the DNA solution. Nine molecular markers were sampled: the barcoding region of mitochondrial Cytochrome Oxidase I (COI); three amplicons of nuclear ribosomal 28S (d-i, f-k, and q-z) (Yang, et al. 2000;

Gibson et al. 2010); one amplicon of alanyl-tRNA synthetase 1 (AATS) (Wiegmann et al. 2011); one amplicon of phosphogluconate dehydrogenase (PGD) (Wiegmann et al. 2011); two amplicons of carbamoylphosphate synthase domain of the rudimentary locus (CAD=CPS), called CAD1 and CAD3 (Moulton and Wiegmann 2004); and 16S rRNA (Gibson et al. 2010). Sequences from the subfamilies Pangoniinae and Tabaninae were available from the Tabanidae PEET Project (Morita et al. submitted; Bayless and Wiegmann 2012). All taxa sampled and genes amplified for each taxon are summarized in Appendix B. Primer pairs used for data acquisition are summarized in the Appendix D. PCR parameters followed reaction protocols specified in Wiegmann et al. (2000), Wiegmann et al. (2011), and Trautwein et al. (2011). PCR products were extracted from agarose gels and purified with the Qiaquick Gel Extraction kit (QIAGEN, Inc.). Big Dye Sequencing kits (Applied Biosystems) were used for sequencing reactions, and sequencing was completed at the North Carolina State University Genomic Sciences Laboratory. Sequences were assembled and edited using Sequencher 4.1 (Gene Codes Corp.).

ALIGNMENT AND PHYLOGENETIC ANALYSES

Alignment was carried out using the software MAFFT (Katoh and Toh 2010). This program calculates multiple sequence alignment rapidly, and is composed of three stages: first, multiple threads make an all-to-all comparison, processing pairwise alignments simultaneously and independently, in order to find an alignment with the highest score based on values assigned to matches or substitutions between similar residues (positive values), infrequent types of substitutions (negative values), and gaps (heavily penalized with negative scores). The second stage is a progressive alignment, which uses a guide tree to calculate group-to-group alignments; and the third stage, called iterative refinement, breaks the alignment into two groups of sequences and realigns them to each other, and if a higher objective score is achieved, this realignment will be kept.

Introns and other positions of ambiguous alignment were removed from the analysis using the least stringent parameters in Gblocks (Castresana 2000). Alignment of protein coding genes according to the amino acid translation was then revised by eye using Mesquite (Maddison & Maddison 2007). All sequences were concatenated in an alignment of 4586 base pairs through Sequence Matrix (Vaidya et al. 2011). In order to address the variations in evolutionary rates and substitution patterns among sites, the data set was partitioned by gene, and for the protein-coding genes, by codon position. The software PartitionFinder (Lanfear et al. 2012) was used to determine the best-fit evolutionary model for all analyses; nine partitions were suggested by this software, as well as an optimal evolutionary model for each partition (table 1). Also, the software PAUP* (Swofford 2002) was used to specify and exclude the third codon positions of the protein coding regions in order to run a second set of both maximum likelihood and Bayesian analyses.

Table 1: Best partition scheme for the concatenated matrix according to PartitionFinder

PARTITION	BEST MODEL	SUBSETS
1	GTR+I+G	16S
2	GTR+I+G	28S
3	GTR+G	AATScodon1, CAD3codon1, PGDcodon1
4	TVM+I+G	AATScodon2, CAD1codon2, CAD3codon2, PGDcodon2
5	TrN+I+G	AATScodon3, CAD1codon3, CAD3codon3, PGDcodon3
6	K81+G	CAD1codon1
7	TrN+I+G	CO1codon1
8	HKY+I+G	CO1codon2
9	HKY+G	CO1codon3

Maximum Likelihood inference was performed using RAxML v.8 (Stamatakis 2014) and GARLI 2.0 (Zwickl 2006). For RAxML, a partitioned analysis was carried out using the GTR+I+G model for all partitions, since this software doesn't apply all the models suggested by PartitionFinder. A total of 250 rapid bootstrap replicates were performed under the frequency-based stopping criterion, and the values were automatically placed on the tree with the best likelihood value. A second RAxML run was performed using the dataset excluding 3rd codon positions, under the same parameters.

The software GARLI supports all the models suggested by PartitionFinder, therefore a partitioned analyses was carried out following the best scheme pictured in table 1 for the 4586 base pairs dataset. For the dataset excluding 3rd codon positions, the GTR+I+G model was applied. Two GARLI runs were performed for each dataset: one with 50 bootstrap replicates and one with non-bootstrap searches in order to obtain the best scoring tree. The program SumTrees (Sukumaran and Holder 2010) was used to draw the bootstrap values on the tree with the best likelihood score found by Garli during the non-bootstrap searches.

Bayesian inference trees were estimated using MrBayes 3.2.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). A partitioned analysis was performed for the 4586 base pair dataset, following the same partition scheme depicted in table 1; however, MrBayes does not support all models recommended by PartitionFinder, therefore the model GTR+I+G was used for partitions 1, 2, 4, 5, 7, and 8, and the model GTR+G was used for partitions 3, 6, and 9. The dataset with 3rd codon positions removed was analyzed under the GTR+I+G model. For both datasets, four simultaneous Markov Chain Monte Carlo (MCMC) chains were set to 20 million generations, with trees being sampled every 10000 generations; the burn-in fraction was set to 0.25, discarding 25% of all trees, and the remaining trees were summarized in a consensus topology with nodes supported by posterior probability values. Stationarity in MCMC runs was assessed using Tracer v.1.6 (Rambaut and Drummond 2009) and convergence was inferred when the standard deviation of the split differences between topologies supported by in-parallel runs was under 0.01.

DIVERGENCE TIME ESTIMATION

The software package BEAST (v2.1.3) (Bouckaert et al. 2014) was used to produce a fossil-calibrated chronogram of estimated clade divergence times; for the sake of computational efficiency, a reduced subset of the concatenated dataset was used. Two species of Athericidae were included in the analyses as outgroup. All genes were individually partitioned and the codon positions for each gene were partitioned to allow positions 1+2 and 3 to be analyzed under unlinked rate parameters. The analysis was performed for 500 million generations sampling every 5000th, using the GTR+I+G substitution model and Yule process of speciation, commencing with a random starting tree. A relaxed lognormal molecular clock model was chosen allowing for rate variation for different gene partitions over time. To calibrate the analysis, the following lognormal priors were used: 1) a minimum age of 44 Ma (m=0.5, sd=0.5) for the clade containing *Rhigioglossa* based on a Bouvieromyiini found in Baltic amber (Trojan 2002); 2) a minimum age of 112 Ma (m=0.5, sd=.05) for the crown group Tabanidae based on *Cratotabanus stonemyomorphus* Martins-Neto & Santos 1994, from the Santana Formation, Brazil; and 3) a minimum age of 142 Ma for the node containing both families Tabanidae and Athericidae, based on the overlapping ages of the fossils *Eotabanoid lordi* Mostovski et al., 2003 from the Durlston Formation, England, and *Laiyangi-tabanus formosus* Zhang, 2012, from the Laiyang Formation, China (Lessard et al. 2013). The resulting trees were subsequently analyzed using LogCombiner to resample every 25000 trees and the burn-in set to 25%, followed by assessment of stationarity in Tracer v1.6 (Rambaut and Drummond 2009) and tree management in TreeAnnotator v2.1.2 to produce a final consensus tree.

RESULTS AND DISCUSSION

TAXON SAMPLING AND ASSEMBLED DATASET

The final alignment resulted in one matrix with 4586 base pairs, and one matrix with third positions of the protein coding genes removed thus a total of 4417 base pairs. Both matrices contain up to six genes (16S, 28S, CO1, AATS, CAD, and PGD) from a total of 77 species: 47 representing 10 genera of the three tribes of Chrysopsinae (*Silvius* and *Chrysops* Meigen, Chrysopsini; *Thaumastocera* and *Orgizomyia* Grunberg, and *Tabanocella* Bigot, Rhinomyzini; *Aegophagamyia* Austen, *Rhigioglossa* Wiedemann, *Lilaea* Walker, *Pseudotabanus* Ricardo, and *Coracella* Philip, Bouvieromyiini), one exemplar of the controversial *Adersia oestroides* (Karsch, 1888), 15 species of 12 genera of Pangoniinae, and 14 species of 14 genera of Tabaninae. 42.54% of the 4586 base pairs dataset contains gap positions. CO1 has the best coverage, followed by 28S, AATS, CAD, 16S, and PGD (Appendix B). Within the datasets, almost no taxa are sampled for all genes, and because of mismatches on gene coverage among taxa, individual gene analyses were considered to be ineffective.

MAXIMUM LIKELIHOOD TOPOLOGIES

RAxML (figures 1 and 2) and GARLI (figures 3 and 4) analyses yielded trees that agree on the composition of the major groupings within Chrysopsinae, however the two different datasets (including and excluding 3rd codon positions) resulted in differences regarding the various relationships among the major lineages.

The tribe Chrysopsini is grouped together in all trees with strong bootstrap support, however it is not recovered as monophyletic due to the presence of *Adersia oestroides* (Karsch, 1888) nested within or as sister to *Silvius*. The genus *Adersia* occurs in the coastal sands of Southern and Eastern Africa to the Arabian Peninsula. The genus type species *Adersia oestroides* was originally described as *Silvius*, but subsequently placed in the controversial subfamily Sepsidinae, along with the Afrotropical genera

Braunsiomyia, and the Neotropical *Sepsis* (Mackerras 1954). These genera are beach dwelling and share a lack of functional mouthparts in adults, but this character is probably convergent (Coscarón 1976). Mackerras revised the placement of *Adersia* when new material became available, stating its stronger affinities with Chrysopsinae (Mackerras 1955). However, he placed *Adersia*, *Braunsiomyia* and *Scepsis* into one clade, suspecting they were a tribe within Chrysopsinae (Mackerras 1955). Mackerras later speculated that *Scepsis* and *Braunsiomyia* could well belong to Pangoniinae, and that *Adersia* was definitely a Chrysopsinae (Dias 1962). In contrast, Dias (1962) gave each genus its own tribe, keeping Braunsiomyiini and Scepsidiini within the Pangoniinae, and separating *Adersia* into its own subfamily (Adersiinae, Dias 1962). The positioning of *Adersia* among other *Silvius* species corroborates Karsch and Mackerras. The cosmopolitan genus *Chrysops* is well supported as monophyletic in all analyses, however the relationships among species remain mostly unresolved. The genus *Silvius* appears as paraphyletic on all analyses; for the analyses with 3rd codon positions removed dataset, it includes *Adersia* and appears as sister group to *Chrysops* with low bootstrap support (figures 1 and 3); on the analyses including 3rd codon positions, *Silvius* appears as intermediate between *Adersia* and *Chrysops* (figures 2 and 4).

The tribe Rhinomyzini is well supported as monophyletic in all maximum likelihood trees, but its position in the tree is not well resolved. Rhinomyzini appear as intermediate between the clades *Pseudotabanus* and (*Coracella* + Tabaninae) with low bootstrap values in the inner node for both GARLI and RAxML analyses with the 3rd codon positions removed dataset (figures 1 and 3). The dataset including 3rd codon positions resulted in a tree where Rhinomyzini appears as intermediate between (*Aegophagamyia + Rhigioglossa*) + (*Chrysopsini + Adersia*) and (*Pseudotabanus +* (*Coracella + Tabaninae*)) (figures 2 and 4). The relationships among the sampled genera of Rhinomyzini are strongly supported by bootstrap values in all maximum likelihood analyses. The monotypic genus *Orgizomyia* from Madagascar is nested within *Tabanocella*, which occurs all over the Afrotropical region; molecular evidence support both genera as one monophyletic group, however more specimens of *Tabanocella* for the molecular dataset, as well as morphological analyses of the type species of both genera, would be necessary to confirm this hypothesis.

The tribe Bouvieromyiini appears as polyphyletic with four distinct lineages. The first clade is a combination of the genera *Aegophagamyia*, from Madagascar, and *Rhigioglossa*, from South Africa, with strong bootstrap support in all analyses, representing a very distinct lineage that could probably be raised to subfamily status. The second clade is represented by the genus *Lilaea*, from Australia, which is also grouped together with strong bootstrap support on all analyses; however, both (*Aegophagamyia + Rhigioglossa*) and *Lilaea* are placed in different positions by the different datasets: the analyses using the excluded 3rd codon positions dataset, for both RAxML and GARLI, places (*Aegophagamyia + Rhigioglossa*) as intermediate between the Pangoniinae and the remaining Chrysopsinae + Tabaninae, and *Lilaea* is recovered as sister to the (Chrysopsini + *Adersia*) (figures 1 and 3); for the dataset including 3rd codon positions on both methods, *Lilaea* appears as intermediate between the Pangoniinae and the remaining Chrysopsini + Tabaninae, and (*Aegophagamyia + Rhigioglossa*) appears as sister to the (Chrysopsini + *Adersia*) (figures 2 and 4). The third lineage of Bouvieromyiini is represented by the genus *Pseudotabanus*, from Australia, and the fourth clade is represented by one species of the Neotropical *Coracella*, which was first described by Philip (1960) as a subgenus of *Mesomyia*. Coscarón (1972), in a more detailed description of the subgenus, stated that the species of *Coracella* are somewhat similar to *Dasybasis* (Tabaninae), and easily distinguished by the hind tibial spur, usually present in Chrysopsinae and absent in Tabaninae. Subsequently, Chainey (1987), stated that the Subcostal vein is bare in all Bouvieromyiini except *Coracella* and the Australian genera *Pseudotabanus* and *Pseudopangonia*, and placed the three groups as subgenera of *Pseudotabanus*, despite the discrepancies in their geographical distribution. All analyses of the dataset excluding the 3rd codon positions place *Pseudotabanus* as intermediate between

((Chrysopsini + *Adersia*) + *Lilaea*) and (*Rhinomyzini* + (*Coracella* + Tabaninae)), while the resulting trees from the dataset including the 3rd codon positions place *Pseudotabanus* as intermediate between the (Rhinomyzini + (*Coracella* + Tabaninae)). Since molecular evidence does not group *Pseudotabanus* and *Coracella* in the same clade, *Coracella* should be reinstated as a genus, disagreeing with Chainey's hypothesis.

Figure 1: RAxML, 3rd codon positions removed, 250 bootstrap replicates, GTR+I+G

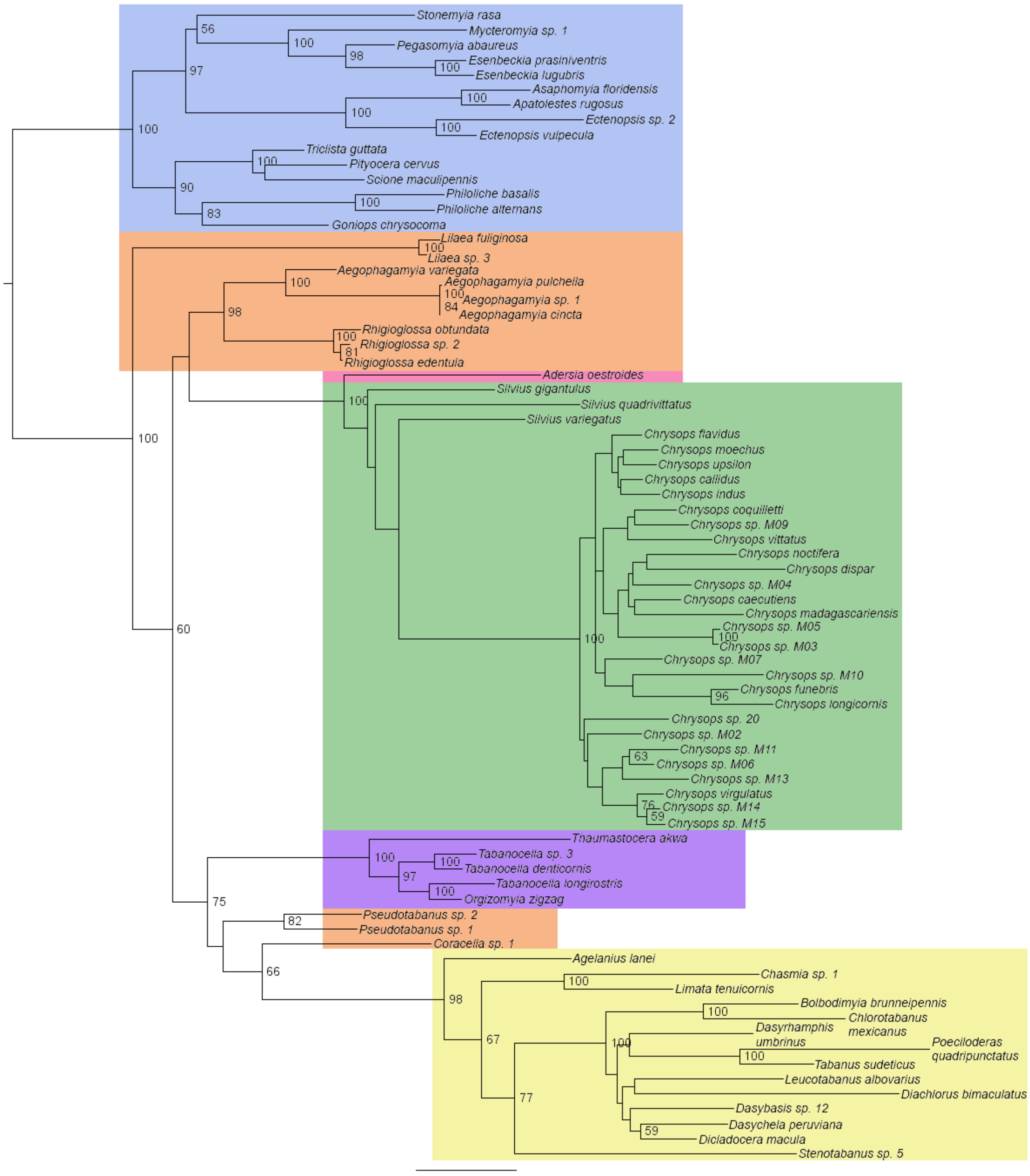

Figure 2: RAxML, 4586 base pairs dataset, 250 bootstrap replicates, GTR+I+G

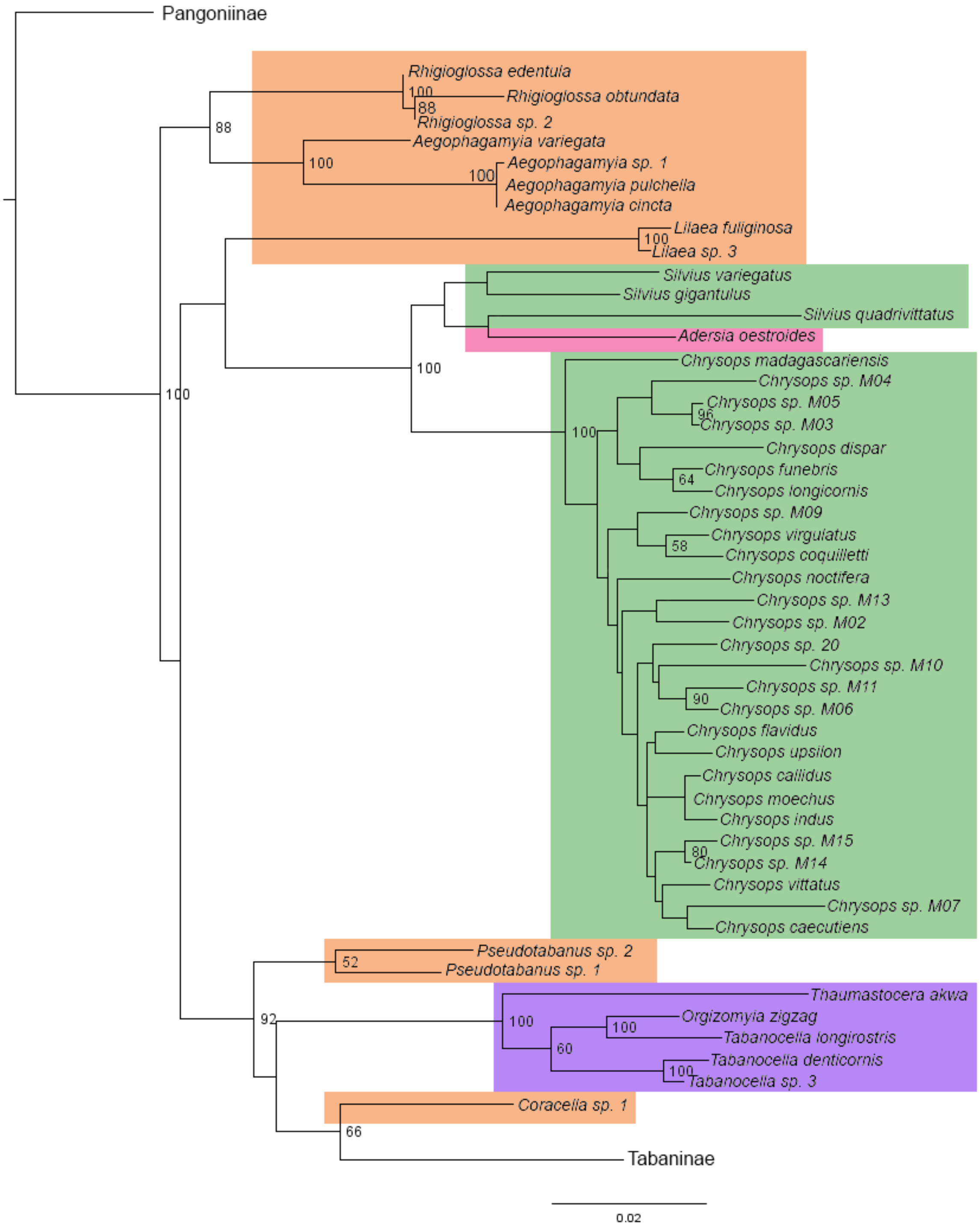

Figure 3: GARLI, 3rd codon positions removed, 50 bootstrap replicates, GTR+I+G

Figure 4: GARLI, 4586 base pairs dataset, 50 bootstrap replicates

Bayesian Inference

The resulting topology from the Bayesian analysis of the dataset with 3rd codon positions removed (figure 5) is approximate to those generated by the maximum likelihood methods. The clade (*Aegophagamyia* + *Rhigioglossa*) appears as intermediate between Pangoniinae and the remaining Chrysopsinae, and *Lilaea* is recovered as sister to the (Chrysopsini + *Adersia*); however, the relationship of *Pseudotabanus*, Rhinomyzini, and (*Coracella* + Tabaninae) remains unresolved – as well as the relationships among species of *Chrysops*.

The Bayesian analysis of the 4586 base pair dataset (figure 6) resulted in a very similar topology to those generated by RAxML and GARLI, including the placement of Rhinomyzini as sister to (*Pseudotabanus* + (*Coracella* + Tabaninae)) with good posterior probability values. The placement of *Lilaea,* (*Aegophagamyia* + *Rhigioglossa*), and (*Adersia* + Chrysopsini) also agrees with those recovered by maximum likelihood methods.

Major findings of these phylogenetic analyses includes confirmation of the paraphyly of Chrysopsinae, a traditional subfamily and once widely accepted among several authors, and the polyphyly of Bouvieromyiini, a tribe mostly grouped by convenience rather than shared morphological traits. The relationships among the four distinct clades of Bouvieromyiini in relation to the other tribes of Chrysopsinae remain uncertain, as well as the sister to the Tabaninae, since all analyses place *Coracella* as sister with low to moderate support and the sampling of Bouvieromyiini genera is limited. The relationships among species remain unsolved, however, this study provides a framework for future systematic revisions of genera and phylogenetic studies of distinct lineages of the Chrysopsinae. In order to address these remaining questions, a more extensive taxon sampling and data collection will be necessary. The increasing availability of next-generation sequencing and phylogenomics will greatly benefit future studies and, ultimately, lead to a better understanding of the relationships of this group.

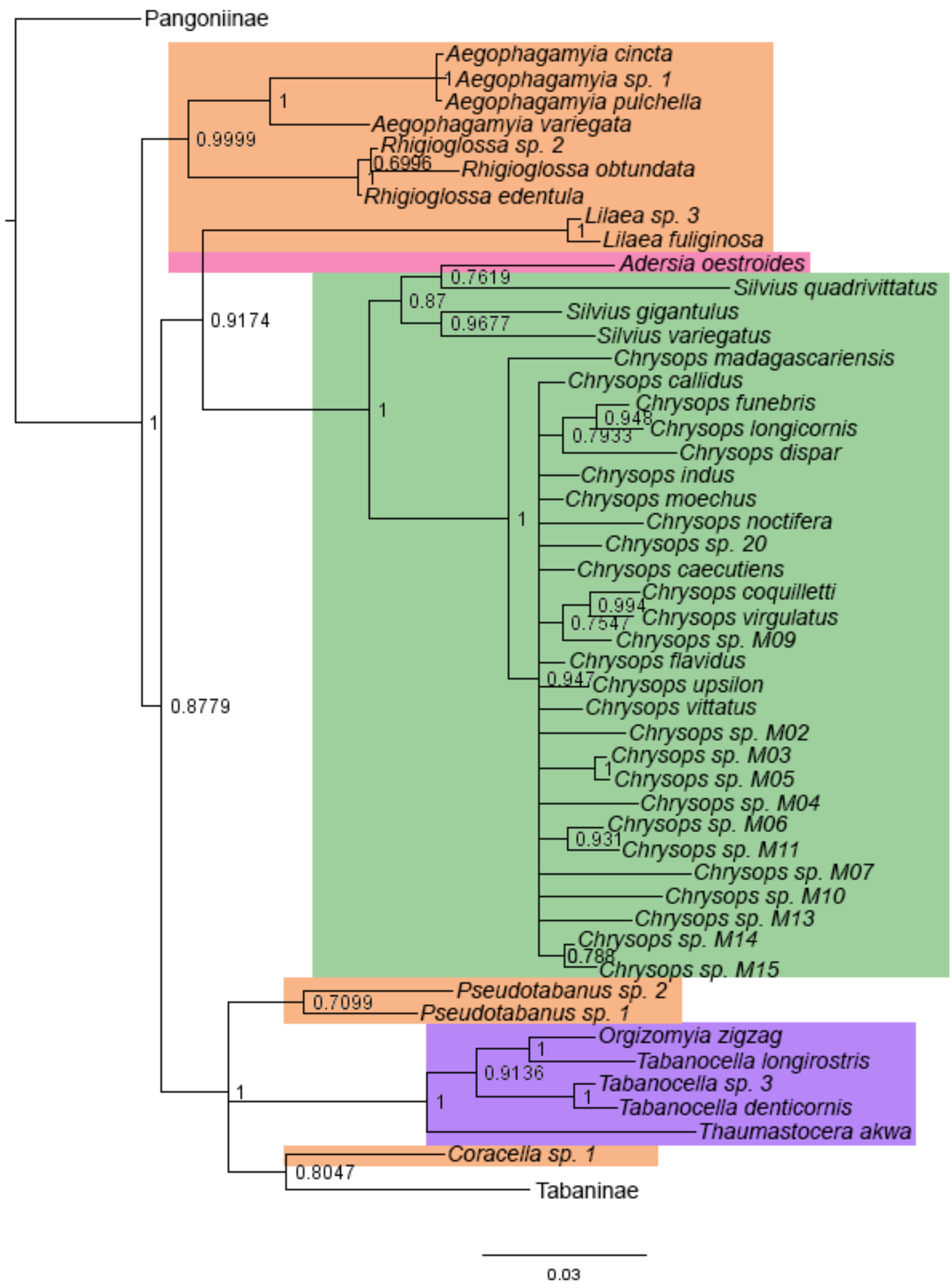

Figure 5: Bayesian inference, 3rd codon positions excluded, 20 million generations, GTR+I+G; posterior probability values displayed on nodes.

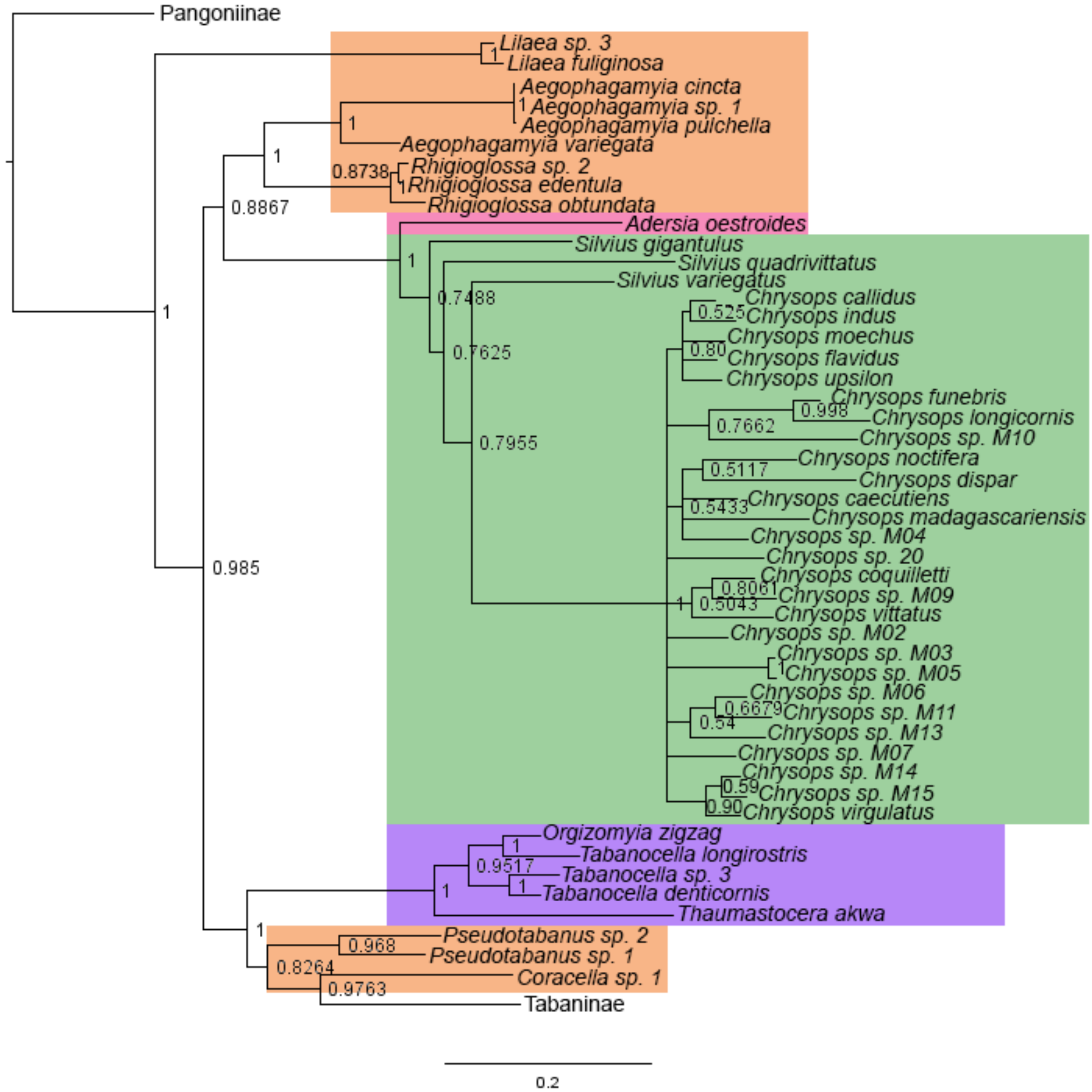

Figure 6: Bayesian inference, 4586 base pairs dataset, 20 million generations, GTR+I+G, posterior probability values displayed on nodes

Divergence Time Estimation

The topology of the divergence time estimation tree (figure 7) agrees in general with the maximum likelihood topologies obtained from the 4586 base pairs dataset. Divergence time estimates show that the major lineages of Tabanidae originated in the early Cretaceous (111.9 Mya), forming a distinct clade of the Pangoniinae as sister to the clade containing the Chrysopsinae + Tabaninae; the Tabaninae diverged from the intermediate lineages known as Chrysopsinae at 69.8 Mya (late Cretaceous/ early Paleogene), and these results are close to those found on a previous study by Lessard et al. (2013). The short internal branches at 103.6Mya, 100.5Mya, and 91.3 Mya suggest a rapid diversification of key lineages of Chrysopsinae. The first lineages to split at 103.6Mya originated the clade containing the Australian genus *Lilaea* and the remaining Chrysopsinae + Tabaninae. The sister clades (*Rhigioglossa* + *Aegophagamyia*) and (Chrysopsini + *Adersia*) separated at 91.3 Mya; the crown of (*Rhigioglossa* + *Aegophagamyia*) began to diversify at 71.4 Mya, and the crown of (Chrysopsini + *Adersia*) at 67.8 Mya. The crown group of Rhinomyzini sampled in this study is a recent group, diversifying at the Oligocene Epoch of the Paleogene (33.5 Mya), however, they diverged from the (*Pseudotabanus* (*Coracella* + Tabaninae)) lineage at the late Cretaceous (87.2 Mya). Finally, the *Pseudotabanus* and (*Coracella* + Tabaninae) also separated in the late Cretaceous (84.7 Mya).

The radiation of the large, cosmopolitan genus *Chrysops* in the Eocene coincides with the divergence of the crown group of Tabaninae, the most species rich subfamily (50Mya), and the diversification of three tribes of Pangoniinae: Pangoniini at 45.1 MY, Philolichini at 58MY, Scionini at 52.7MY (Lessard et al. 2013). Also occurring in this epoch was a major radiation of several orders of modern large mammals such as the Perissodactyla, which includes the horses. However, during the Cretaceous, when the major lineages of horse flies diverged, there were no representatives of large mammals similar to the prominent mammals of today: the majority of mammals were small in size (e.g. the size of shrews, rats, and mice)(Kemp 2005), which might suggest that horse flies switched host preferences from small to large mammals, or even that the early lineages fed on other vertebrates such as reptiles or birds – a behavior that is recorded for a few species of modern horse flies (Henriques et al. 2000, Ferreira and Rafael 2004).

These findings allow us to delineate the diversification dates of major lineages of horse flies, and although it is not sufficient to make inferences on the global radiation patterns of Chrysopsinae yet, these results provide a backbone for future work. Additional molecular markers as well as a more thorough taxon sample will be necessary to provide a more refined evaluation of the history of diversification of these flies.

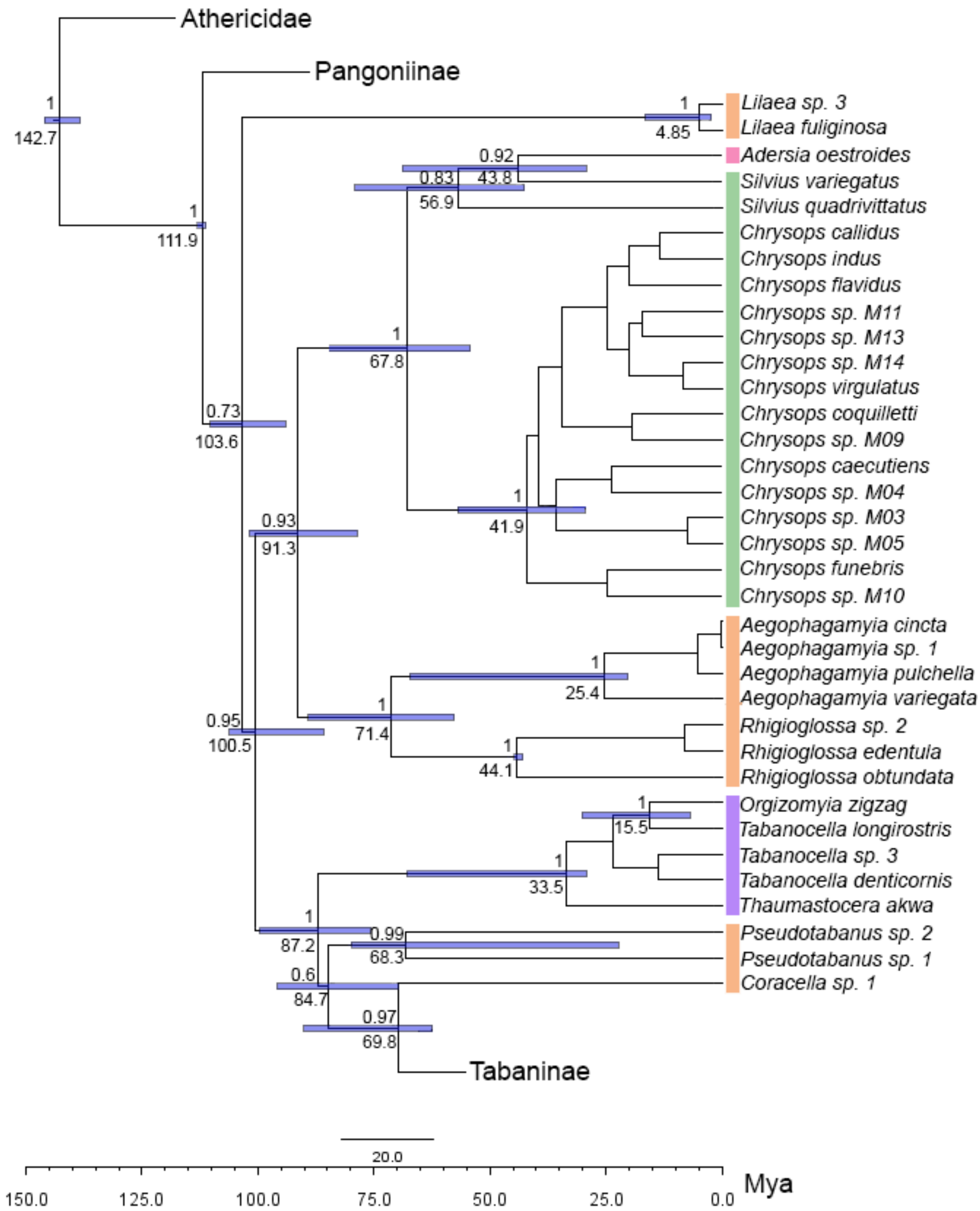

Figure 7: Bayesian Divergence Time Estimation of the Tabanidae. GTR+I+G, 500 million generations

CONCLUSIONS

Our results corroborate the previous hypothesis that the Chrysopsinae is not a natural group, and therefore, should not be considered a subfamily in current classifications of the family Tabanidae. Although relationships among the major lineages of Chrysopsinae remain somewhat inconclusive, this study is the first molecular study specifically aimed at expanding our knowledge of the phylogenetic history of deer flies, representing a starting point to redefine this well-known, but taxonomically difficult, group. Only the tribe Rhinomyzini emerges as monophyletic, however its placement in relation to other lineages of deer flies is also not certain. Bouvieromyiini is paraphyletic and demands extensive revisionary work in order to delineate stable groups and their position in the phylogenetic tree. Chrysopsini should include the former Sepsidinae genus *Adersia*, and the cosmopolitan genus *Silvius* also needs revisionary attention. Future work should expand taxon coverage, including as many genera of former Chrysopsinae as possible, as well as take advantage of the increasing availability of next-generation sequencing technologies and phylogenomics to achieve these goals.

REFERENCES

Bayless, K.M. and Wiegmann, B.M. (2012). *Evolutionary Studies in Tabanomorpha (Diptera) with New Classifications of Xylophagidae and Tabaninae.* Unpublished manuscript. Department of Entomology, North Carolina State University, Raleigh, North Carolina.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T.G., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J. (2014). BEAST2: A software platform for Bayesian evolutionary analysis. *PLOS Computational Biology 10*(4): e1003537.

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

Chainey, J.E. (1987). Afrotropical Tabanidae (Diptera): the genus Rhigioglossa Wiedemann, 1828 (including *Mesomyia* Macquart, 1850, as a subgenus). *Annals of the Natal Museum 28*: 137-159.

Chainey, J.E. (1993). Horse–flies, deer-flies and clegs (Tabanidae). In: Lane PR, Crosskey RW. (Eds) Medical Insects and Arachnids. Chapman & Hall, London, 706pp.

Coscarón, S. (1972). Contribución al conocimiento de los Tabanidae neotropicales: I. El subgénero Mesomyia (Coracella) Philip (Diptera-Tabanidae). *IDESIA 2*: 87-97.

Coscarón, S. (1976). Contribución al conocimiento de los Tabanidae neotropicales II: los Pangoniini del Sur de Sudamérica y datos sobre la tribu Scepsidini. *Revista del Museo de La Plata. Sección Zoología 12* (114): 75-116.

Dias, J.A.T.S. (1962). Tentativa de uma nova classificação dos insectos da família Tabanidae Loew, 1860. *Memórias e estudos do Museu Zoológico da Universidade de Coimbra 277*: 1-13.

Fairchild, G.B. (1969). Notes on Neotropical Tabanidae XII. Classification and distribution, with keys to genera and subgenera. *Arquivos de Zoologia de São Paulo 17*: 199-255.

Ferreira, R.L.M. and Rafael, J.A. (2004). Activities of tabanids (Diptera, Tabanidae) attacking domestic duck *Cairina moschata* (Linnaeus) (Aves, Anatidae), introduced in a forest area in the Central Amazon, Manaus, Brazil. *Revista Brasileira de Entomologia 48*(2): 283-286.

Gibson, J.F., Skevington, J.H., Kelso, S. (2010). Placement of Conopidae (Diptera) within Schizophora based on mtDNA and nrDNA gene regions. *Molecular Phylogenetics and Evolution, 56*(1): 91-103.

Henriques, A. L.; Ferreira, R. L. M.; Vidal, J. F., and Rafael, J. A. (2000). *Betrequia ocellata* Oldroyd (Diptera, Tabanidae, Rhinomyzini) blood feeding on *Caiman crocodilus* (Linnaeus) (Crocodylia, Alligatoridae) in Manaus, Brazil. *Revista Brasileira de Zoologia 17*(3): 609–613

Huelsenbeck, J. P. and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics 17*:754-755.

Kemp, T. S. (2005). *The Origin and Evolution of Mammals*. Oxford University Press, Oxford, U.K. x + 331 pp.

Krinsky, W. L. (1976). Animal disease agents transmitted by horse flies and deer flies (Diptera: Tabanidae). *Journal of Medical Entomology 13*: 225–275

Katoh, T. & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics 26*(15): 1899-1900.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution 29*(6): 1695-1701.

Lessard, B. D. and Yeates, D. K. (2012). Seven new Australian species of the Southern hemisphere horse fly subgenus *Scaptia (Pseudoscione)* (Diptera: Tabanidae), including descriptions and a revised key. *Journal of Medical Entomology 49*(6): 1206–1225.

Mackerras, I. M. (1954). The classification and distribution of Tabanidae (Diptera) I. General Review. *Australian Journal of Zoology 2*: 431–454.

Mackerras, I. M. (1955). The classification and distribution of Tabanidae (Diptera) III. Subfamilies Sepsidinae and Chrysopsinae. *Australian Journal of Zoology 3*: 583–633.

Maddison, W. P. and Maddison, D.R. (2007). Mesquite: a modular system for evolutionary analysis. Version 2.0 http://mesquiteproject.org

Martins-Neto, R.G, Santos, J.C.K. (1994). Um novo gênero e uma nova espécie de Mutuca (Insecta, Diptera, Tabanidae) da Formação Santana (Cretáceo Inferior), Bacia do Araripe, Nordeste do Brasil. *Acta Geologica Leopoldensia 39*: 289-297.

Morita, S.I. (2008). A phylogeny of long-tongued horse flies (*Philoliche*, Diptera: Tabanidae) with the first cladistic evaluation of higher relationships within the family. *Invertebrate Systematics 22*(3): 311–327.

Morita, S.I., Bayless, K.M., Yeates, D.K., Wiegmann, B.M. Molecular phylogeny of the horse flies: a framework for renewing tabanid taxonomy. *Systematic Entomology*, in press.

Mostovski, M.B., Jarzembowski, E.A., Coram, R.A. (2003). Horseflies and athericids (Diptera: Tabanidae, Athericidae) from the lower Cretaceous of England and Transbaikalia. *Paleontological Journal 37*: 162–169.

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

Oldroyd, H. (1957). The Horse-flies (Diptera: Tabanidae) of the Ethiopian Region III: Chrysopsinae, Scepsidinae, and Pangoniinae and a Revised Classification. British Museum, London. 489 pp.

Pape, T., & Thompson, F. C. (2012). Systema dipterorum (version 2.0, Jan 2011). In: Bisby, F., Roskov, Y., Culham, A., Orrell, T., Nicolson, D., Paglinawan, L., Bailly, N., Kirk, P., Bourgoin, T., Baillargeon, G. (Eds.), Species 2000 & ITIS Catalogue of Life. Digital resource at www.catalogueoflife.org/col/. Species 2000, Reading, UK. Accessed (05.13.13)

Philip, C. B. (1960). Further records of Neotropical Tabanidae (Diptera) mostly from Peru. *Proceedings of the California Academy of Sciences, Fourth Series 31*(3): 69-102.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*:1572-1574.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics 30*(9):1312-1313.

Stuckenberg, B. R. (2001). Pruning the tree: a critical review of classifications of the Homeodactyla (Diptera, Brachycera), with new perspectives and an alternative classification. *Studia Dipterologica 8*(1), 3–41.

Sukumaran, J. and Mark T. Holder. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics 26*: 1569-1571.

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

Trautwein, M.D., Wiegmann, B.M., Beutel, R., Kjer, K.M., & Yeates, D.K. (2012). Advances in Insect Phylogeny at the Dawn of the Postgenomic Era. *Annual Review of Entomology 57*: 449-68.

Trojan, P. (2002). First discovery of Bouvieromyiini (Diptera: Tabanidae: Chysopsinae) in Baltic amber. *Annales Zoologici (Warsaw) 52*: 257-270

Usher, P. J. (1972). A review of the South African horsefly fauna (Diptera: Tabanidae). *Annals of the Natal Museum 21*: 459–507

Vaidya, G., Lohman, D. J. and Meier, R. (2011), SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics 27*: 171–180.

Wiegmann, B.M., Tsaur, S.C., Webb, D.W., Yeates, D. K., and Cassel, B. K. (2000). Monophyly and relationships of the Tabanomorpha (Diptera: Brachycera) based on 28S ribosomal gene sequences. *Annals of the Entomological Society of America 93*(5): 1031–1038.

Wiegmann, B.M., Trautwein, M., Winkler, I.S., Barr, N.B., Kim, J.W., Lambkin, C., … Yeates, D.K. (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences* March 14, 2011.

Yang, L., Wiegmann, B.M., Yeates, D.K., Irwin, M.E. (2000). Higher-Level Phylogeny of the Therevidae (Diptera: Insecta) Based on 28S Ribosomal and Elongation Factor-1α Gene Sequences. *Molecular Phylogenetics and Evolution 15*(3): 440-451.

Zhang, J. (2012). New horseflies and water snipe-flies (Diptera: Tabanidae and Athericidae) from the lower Cretaceous of China. *Cretaceous Research 36*: 1–5.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

CHAPTER 2: Anchored Phylogenomics of Chrysopsinae

INTRODUCTION

High-throughput sequencing technologies are rapidly becoming highly useful for researchers in the fields of molecular systematics and phylogenetics, due to the possibility of obtaining large amounts of molecular data, including from non-model organisms, in a cost- and time-effective way (Lemmon et al. 2012, Lemmon and Lemmon 2013, McCormack et al. 2013). Previous phylogenetic work has relied on labor intensive techniques such as targeted polymerase chain reaction (PCR), which allows the generation of data from small numbers of genes at a time, frequently resulting in unresolved branches at different phylogenetic levels (Lemmon et al. 2012). While many major findings resulted from PCR-based molecular systematics in the last 25 years (Trautwein et al 2012), deficiencies in sampling for both taxa and genes, unpredictability in information content and evolutionary rate for amplifiable genes, and difficulties in assigning branch lengths for rapid radiations of species, have all severely limited the resolving power of PCR-based approaches for many major higher-level phylogenetic questions. The six gene study of Chapter 1, using the aforementioned approach proved largely inconclusive in reconstructing relationships among the major lineages of the horse fly subfamily Chrysopsinae. New data sets gained from large numbers of orthologous genes identified using next-generation sequencing and comparative bioinformatics approaches show great promise in providing an enormous new source of information that can be used to help resolve the most difficult nodes in the fly tree of life.

One emerging technique using bioinformatics and large-scale next-generation sequencing is called Anchored Enrichment (Lemmon et al. 2012). This is a promising approach for both shallow and deep level phylogenetic studies of non-model groups, without the need of a reference genome. In this method, specific genomic regions are captured prior to sequencing; capture probes for the target regions are designed from known sequences, aiming highly conserved anchor regions with rapidly evolving adjacent regions to provide sufficient phylogenetic information (Lemmon et al. 2012).

The probes are then mixed with genomic DNA either on an array or in solution, and only DNA that successfully hybridizes to these probes is sequenced using high-throughput technology (McCormack et al. 2013, Lemmon and Lemmon 2013, Hedtke et al. 2013). One of the major advantages of anchored enrichment is the ability to use material preserved in ethanol, frozen tissues, and even old DNA and partially degraded samples because the probes are short sequences and many small copies of the bound region and its flanking sequence can be assembled from high-coverage ('deep') sequencing of the bound fragments (Lemmon and Lemmon 2013, Hedtke et al. 2013) (figure 1).

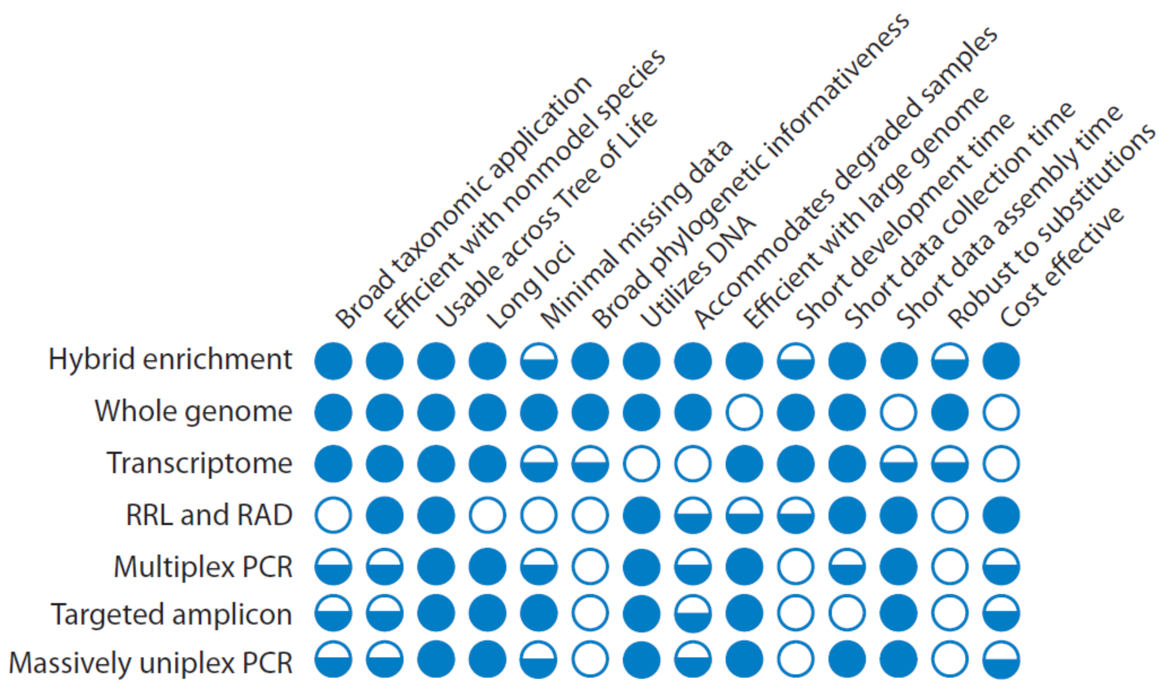

Figure 1: Comparison of attributes of genomic partitioning approaches for phylogenetic data collection (extracted from Lemmon and Lemmon 2013)

In this chapter, maximum likelihood phylogenetic inference of the former subfamily Chrysopsinae was performed using molecular data generated by a custom set of anchored enrichment probes. With this analysis we intend to achieve a better resolution of the higher-level phylogenetic relationships of deer files from hundreds of additional gene regions obtained from probe capture and next-generation sequencing technology.

MATERIAL AND METHODS

Nuclear genomic DNA was extracted using the DNeasy DNA extraction kit (QIAGEN, Inc.). The standard protocol was altered by extending the amount of time the specimen was in proteinase K solution to 2 days to allow more cells to be broken down. The final elution was reduced to 30 µL to avoid diluting the DNA solution. DNA templates were sent to the Center for Anchored Phylogenomics at the Florida State University, where our collaborators conducted the locus selection and probe design, enrichment, and sequencing.

Gene targets were identified and designed through collaboration with A. Lemmon (FSU). To identify orthologous gene target regions, we compared genes from eight dipteran transcriptomes and five complete genomes from a broad range of fly families and including one Tabanidae (*Chrysops flavidus*). These were used to identify potential target sequences useful across the order Diptera. Candidate probes (baits) were assessed for target specificity by comparing each potential probe against the entire *Drosophila melanogaster* genome using BLASTN (Altschul et al. 1997). Probes were selected that hybridize only to the coding sequence of interest (the ''target exon''), and thus could be successful at capturing that exon prior to sequencing. Probes that hit multiple regions of the reference genome with a high bitscore are excluded because these sequences may capture non-homologous gene regions. Probes with putative single-copy orthologs in a majority of the sampled transcriptomes are aligned across all taxa and areas of high conservation across the set are retained. Probes are also screened for GC content; only candidate baits with a GC content of 40–60% were retained. The final probe sequences are from 121- 1497 bp long, with an avg. length of 588 bp. The final set of probes for use across Diptera resulted 650 candidate exon sequences that will target gene regions 521-1897bp in length, with an avg. length of 988 bp. Across all of the dipteran taxa used in the design these probes should capture approximately 539,666 bp.

We used a protocol modified from Meyer and Kircher (2010) to construct an indexed DNA library for each species. The Illumina adapters included index sequences and these were sequenced in a separate indexing read. The Library was standardized to contain 300-400 bp insert sizes. The library was enriched using an Agilent Custom SureSelect kit containing a single pool of all Dipteran probes. Hybridization protocols followed the manufacturers recommended procedure and detailed in Lemmon et al. (2012). High throughput sequencing was performed on the Illumina HiSeq 2000 sequencing system (all three multispecies pools were combined in a single lane). Paired-end 100-bp reads were produced on the HiSeq 2000. Sequencing included an 8-bp indexing read.

Raw sequence processing, assembly, and alignment were carried out as detailed in Lemmon et al. (2012). After assembly, loci were retained as captured only if there were 60 or more total reads for that species, to account for a commonly high number of misindexing errors (Kircher et al. 2012; Lemmon et al. 2012). Assemblies were also processed through an automated refinement step to remove obvious errors and adjusted manually in Geneious© Pro v6.2 (Biomatters Limited). For each locus, consensus sequences were aligned using MUSCLE (Edgar 2004), with default parameters as implemented in Geneious© Pro v6.2.1. Alignments were inspected by 'eye', and ambiguous regions or missing sites (i.e., "N" or "-") were excluded from phylogenetic data sets, thus there was no missing data within a locus in the final alignment. Because of high variability outside the probe region in many loci, the entire probe region and as much of the adjacent regions on both sides as could be reliably aligned were included in a single character block. These character blocks were concatenated in a single data set for phylogenetic analysis.

The software jModelTest (Darriba et al. 2012, Guindon & Gascuel 2003) was used to determine the most appropriate model of nucleotide substitution for the full concatenated alignment. Maximum Likelihood inference was performed using RAxML v.8 (Stamatakis 2014) and GARLI 2.0 (Zwickl 2006), both applying the GTR+I+G model as determined by jModelTest. For RAxML, a total of 200 rapid bootstrap replicates were performed under the frequency-based stopping criteria, and the values were automatically drawn in the tree with the best likelihood value.

Two GARLI runs were performed: one with 100 bootstrap replicates and one with non-bootstrap searches in order to obtain the best scoring tree. The program SumTrees (Sukumaran and Holder 2010) was used to draw the bootstrap values on the tree with the best likelihood score found by GARLI during the non-bootstrap searches.

RESULTS AND DISCUSSION

The final alignment resulted in a 106,647 bp dataset for 19 taxa, for three subfamilies of Tabanidae: one *Esenbeckia* Rondani and one *Ectenopsis* Macquart (Pangoniinae); one *Poeciloderas* Lutz (Tabaninae); and for the family Chrysopsinae, we sampled: one *Neochrysops* Szilady, two *Silvius* and three *Chrysops* (Chrysopsini); one *Rhigioglossa*, one *Merycomyia* Hine, two *Pseudotabanus*, and two *Lilaea* (Bouvieromyiini); and one *Orgizomyia*, two *Tabanocella*, and one *Thaumastocera* (Rhinomyziini). The proportion of gaps and completely undetermined characters in the alignment was 19.03%. The concatenated dataset delivered by our collaborators did not include a partition file, and for this reason we were unable to assess how many and which molecular markers compose the dataset.

Both RAxML and GARLI analyses resulted in the same tree topology (figures 2 and 3) yielding good bootstrap support on all nodes, although, this topology is different from those generated by standard PCR data (chapter 1). For the anchored enrichment dataset, the genera *Chrysops* and *Silvius* both appear as monophyletic with 100 percent bootstrap support. The Chrysopsini species *Neochrysops globosus* was included in this analyses, and it appears as intermediate between the Pangoniinae and the remaining (Chrysopsinae + Tabaninae) with a very long branch; one possible explanation is that this template might be compromised by contamination, since the DNA was extracted from a pinned specimen from the National Museum of Natural History (NMNH), and I was unable to get any sequences from this template via standard PCR techniques.

The Bouvieromyiini appears split in different lineages across the tree, in concordance with the results found on Chapter 1, however, the position of these lineages is the main difference between the topology generated from anchored enrichment and those from standard PCR. The clade (*Rhigioglossa* + *Merycomyia)* appears as sister to the (*Chrysops* + *Silvius*) with 100 percent bootstrap support; *Rhigioglossa* also appeared as sister to the Chrysopsini on the trees generated by standard PCR datasets including 3rd codon positions, however, *Merycomyia* was not included in the previous study – the DNA for this species was also extracted from a pinned specimen from the NMNH, and also didn't amplify via standard PCR methods, but in this case, it doesn't seem to be compromised by contamination. The clades (*Chrysops* + *Silvius*) + (*Rhigioglossa* + *Merycomyia*) are combined in a clade with moderate bootstrap support (57 for RAxML and 63 for GARLI). The second Bouvieromyiini clade is *Pseudotabanus*, which position was close to the Rhinomyzini in the trees generated by both datasets used in the previous study, but here is recovered as sister to the Rhinomyzini on a separate clade with strong bootstrap support. The third and last clade of Bouvieromyiini is formed by the Australian genus *Lilaea*, here positioned as sister to the Tabaninae with good bootstrap support. The position of *Lilaea* in the previous chapter depended on the inclusion or exclusion or the 3rd codon positions in the dataset; however, in either cases, it was never recovered as sister to the Tabaninae as seen in the resulting tree from the anchored enrichment dataset.

Finally, the Rhinomyzini is recovered as monophyletic with strong bootstrap support on all nodes of ((*Tabanocella* + *Orgizomyia*) + *Thaumastocera*), in concordance with the results from standard PCR methods in chapter 1.

The amount of data is a logical explanation for the differences on the trees generated here and in the previous study, since the dataset generated by Anchored Enrichment is over 20 times larger than the one generated by standard molecular techniques. However, an increased amount of data also means an increased complexity, which can result in stochastic error (lack of precision, e.g. when the data doesn't contain sufficient phylogenetic information) and systematic error (lack of accuracy). Systematic error usually is a result of incorrect model selection, for the larger the dataset, the more difficult it is to choose accordingly, which might result in strongly supported yet inaccurate phylogenetic reconstruction (Weisrock 2012, Lemmon and Lemmon 2013). Another possible source of systematic error is the concatenation of conflicting loci. Simulation and empirical studies have shown that concatenating sequences that present high level of discordance on their gene trees also leads to strongly supported but incorrect phylogenetic inference (Weisrock et al. 2012). The Anchored Enrichment approach is very promising for phylogenetic inference, and future work should not only increase the taxon sampling, but also focus on the identification of the molecular markers that compose the assembled matrix in order to refine the phylogenetic analyses (e.g. using individual gene tree methods, partitioned dataset with multi-model analyses) with more confidence in the dataset generated.

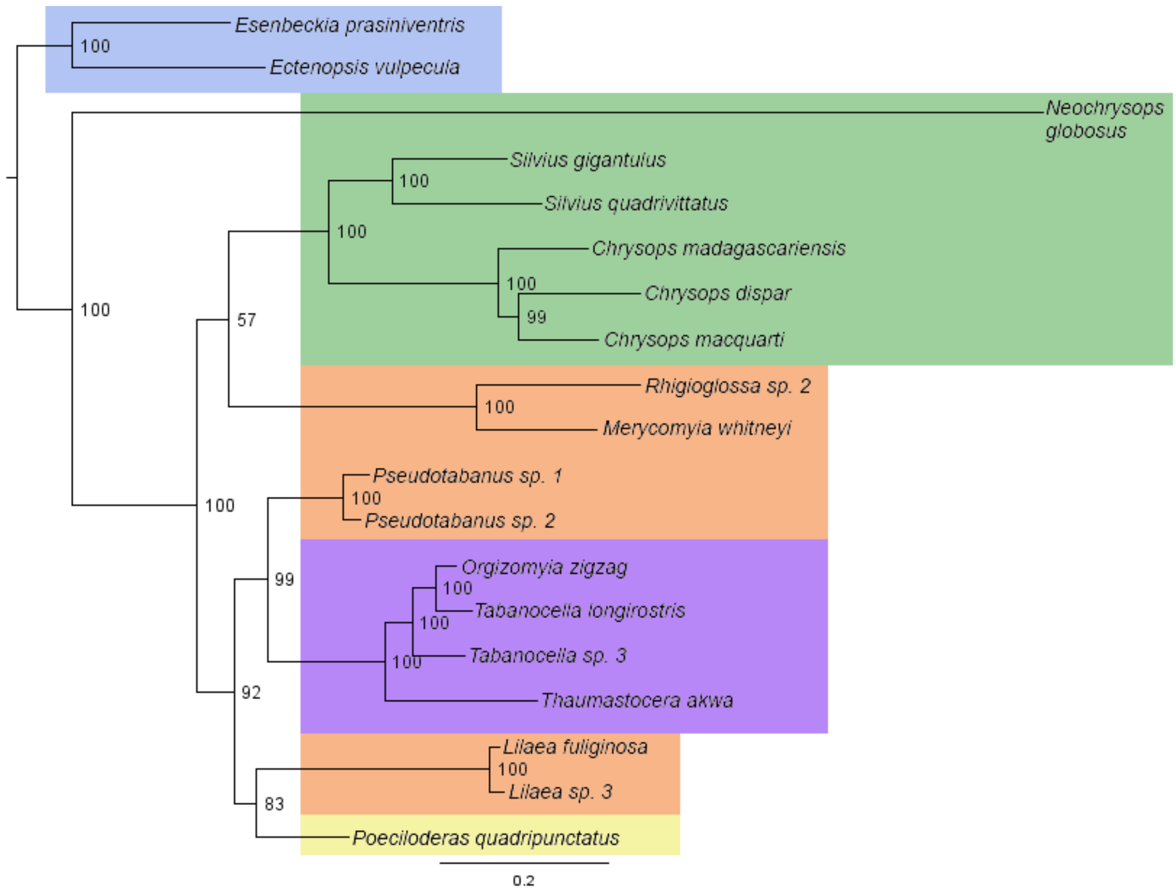

Figure 2: RAxML, GTR+I+G, 200 bootstrap replicates

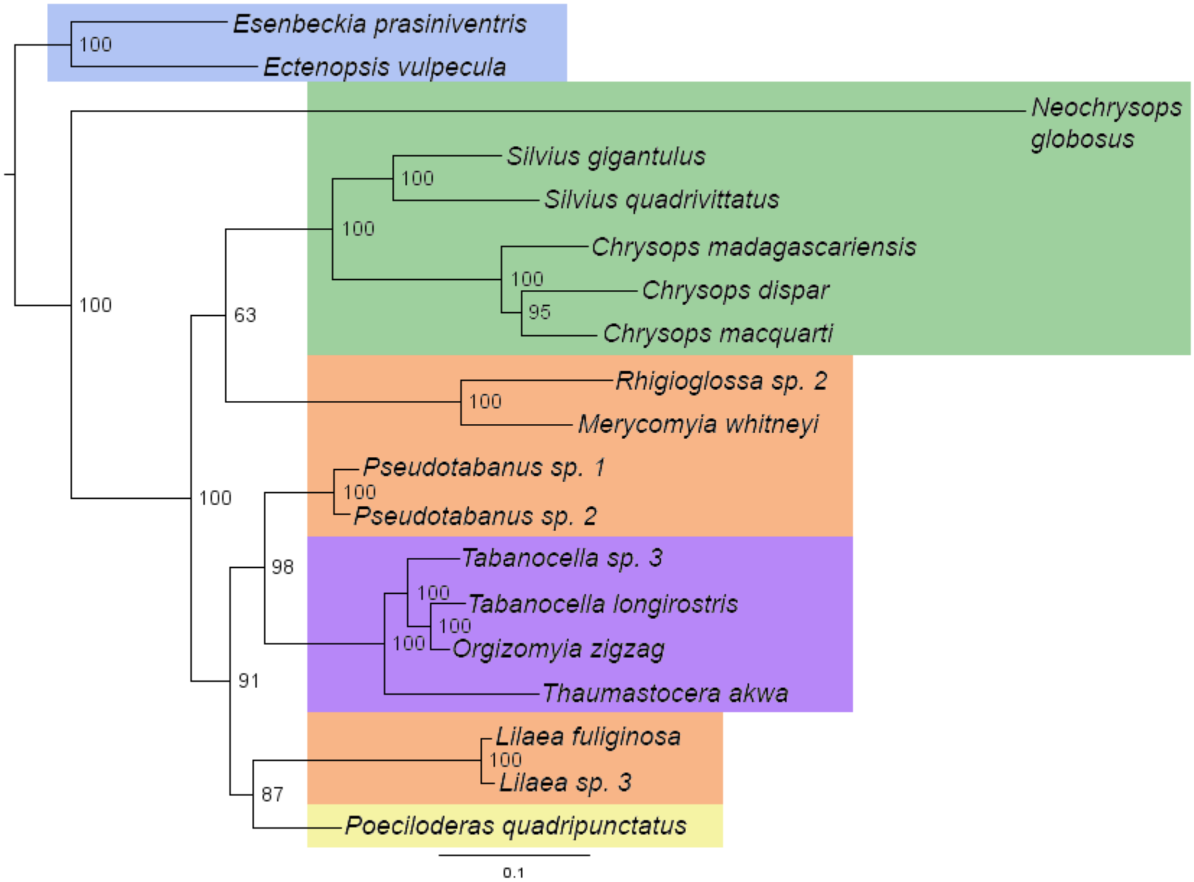

Figure 3: GARLI, GTR+I+G, 100 bootstrap replicates

CONCLUSIONS

In this study, a novel next-generation sequencing technique was used for the first time to generate a phylogeny of horse flies. Despite the limited number of taxa, it yields consistent results and represents the first step towards the delineation of a new classification based on stable, monophyletic groups. This approach is very promising for phylogenetic inference, and since the probes for the former Chrysopsinae have already been developed and tested, it is expected to greatly reduce the time needed for data collection, and to deliver high-quality genomic data for a fraction of the cost required by other methods.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25(17)*: 3389–3402.

Darriba, D., Taboada, G.L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods 9*(8): 772.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*(5): 1792-97.

Guindon, S. & Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology 52*: 696-704.

Hedtke, S.M., Morgan, M.J., Cannatella, D.C., & Hillis, D.M. (2013). Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time. *PLOS ONE 8*(7): e67908.

Kircher, M., Sawyer, S., Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research 40(1)*: e3.

Lemmon, A.R., Emme, S.A., & Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology 61*: 727–744.

Lemmon, E.M. and Lemmon, A.R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Reviews of Ecology, Evolution and Systematics 44*: 99–121.

McCormack, J.E., Hird, S.M., Zellmer A.J., Carstens, B.C., Brumfield, R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution 66*: 526–38.

Meyer, M. and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols 2010*: doi:10.1101/pdb.prot5448.

Rambaut, A. and Drummond, A.J. (2009). Tracer v1.6. Edinburgh (United Kingdom): Institute of Evolutionary Biology, University of Edinburgh. Available from: http://beast.bio.ed.ac.uk/software/tracer/

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics 30*(9):1312-1313.

Sukumaran, J. and Holder, M.T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics 26*: 1569-1571.

Weisrock, D.W., Smith, S.D., Chan, L.M., Biebouw, K., Kappeler, P.M., and Yoder, A.D. (2012). Mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. Molecular Biology and Evolution 29: 1615–1630.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

CHAPTER 3: Phylogenomics of Tabanomorpha: the Agalma Workflow

INTRODUCTION

Tabanomorpha is a brachycerous infraorder composed of 5537 species in nine families; nearly all are predators as larvae and most adults feed on nectar and pollen, but several also feed on blood of vertebrates (Wiegmann et al. 2000; Kerr 2010). The nine families that currently compose Tabanomorpha are: Athericidae Stuckenberg 1973; Austroleptidae Nagatomi 1982; Bolbomyiidae Rohdendorf & Rasnitsyn 1980; Oreoleptidae Zloty, Sinclair & Pritchard 2005; Pelecorhynchidae Mackerras & Fuller 1942; Rhagionidae Samouelle 1819; Tabanidae Samouelle 1819; Vermileonidae Nagatomi 1975; and Xylophagidae Stephens 1829. Tabanomorpha is divided into two superfamilies: Tabanoidea, including the families Tabanidae, Athericidae, Oreoleptidae and Pelecorhynchidae; and all other families compose Rhagionoidea (Kerr 2010; Bayless and Wiegmann 2012). Both Tabanomorpha and Tabanoidea are recovered as monophyletic based on molecular evidence (Wiegmann et al. 2000, Kerr 2010, Wiegmann et al. 2011), as well as the families included in this study: Rhagionidae (Kerr 2010), Pelecorhynchidae (Wiegmann et al. 2000), Tabanidae (Wiegmann et al. 2000; Mackerras 1954; Stuckenberg 2001), and Xylophagidae (Bayless and Wiegmann 2012).

A growing number of phylogenetic studies are relying on transcriptome data as a cost-effective alternative to whole genome sequencing (Lemmon and Lemmon 2013). In this approach, called RNA-Seq, RNA extraction is performed from specific tissues of the target organism. RNA is then reverse transcribed into cDNA, and each molecule is sequenced with high-throughput technology; the resulting sequence reads can either be aligned with the reference genome or transcripts or assembled using *de novo* strategies – for this reason, RNA-Seq is attractive for non-model organisms without a sequenced genome (Lemmon and Lemmon 2013, Wang et al. 2009). Other advantages of this method are the production of large amounts of informative data for different phylogenetic levels from small amounts of RNA sample, and the generation of robust phylogenetic inferences resulting from analyses of very large data matrices from reassembled transcribed gene fragments. This method is proving to be highly informative of organismal phylogeny even if constructed from few sequence reads (Wang et al. 2009, Hittinger et al. 2010, Lemmon and Lemmon 2013; Breinhold and Kawahara 2013). However, one drawback of this approach is the complexity of the analyses, with numerous steps that require different and often non-integrated tools, which are difficult to implement and to describe in a precise manner, which may cause problems in interpretation or reproducing results (Dunn et al. 2013). An automated tool called Agalma was developed by Dunn et al. 2013 to construct matrices for phylogenetic analyses from raw transcriptome data or externally produced assemblies; if raw transcriptome data is provided, Agalma may be used to produce annotated assemblies, an aligned gene sequence matrix, a preliminary phylogenetic tree, as well as a detailed report of the analyses steps and the final results.

In this chapter, I use transcriptome data of 10 species of brachyceran flies with previously known phylogenetic relationships to test Agalma as a tool for generating alignments of homologous genes and building species trees.

MATERIAL AND METHODS

Assembled transcriptome data was obtained through the collaboration of Dr. Brian Wiegmann with 1Kite, a project aiming to sequence 1,000 insect transcriptomes (1K Insect Transcriptome Evolution). Ten taxa were selected from the 1Kite species list: three species of Rhagionidae (*Symphoromyia sp.*, *Chrysopilus quadratus*, and *Rhagio sp.*); two species of Pelecorhynchidae (*Pelecorhynchus rubidus* and *Pelecorhynchus fulvus*); three species of Tabanidae (*Scaptia jacksoniensis*, *Chrysops flavidus*, and *Haematopota fluvialis*), and two species of Xylophagidae as outgroup (*Dialysis sp.* and *Xylophagus abdominalis*).

The assemblies were directly catalogued and loaded into the Agalma database, and then the *homologize* pipeline was executed, which uses an all-by-all tblastx search to identify homologous sequences across the loaded assemblies; next, the *multalign* pipeline was used to apply sampling and length filters to each cluster of homologous sequences, followed by multiple sequence alignment using MACSE (Ranwez et al. 2011) and removal of poorly aligned positions using Gblocks (Castresana 2000). The *genetree* pipeline used RAxML (Stamatakis 2014) to build a maximum likelihood tree for each set of homologous sequences; then, the *treeprune* pipeline, which identifies orthologs based on the topology of gene phylogeny, using DendroPy (Sukumaran and Holder 2010) to prune each gene tree into maximally inclusive subtrees with only one sequence per taxon; each subtree is considered a set of putative orthologs, and were then re-entered as clusters into Agalma's database, and the *supermatrix* pipeline was executed to concatenate the multiple alignments into a supermatrix (table 1); and finally, the *speciestree* pipeline was used to build a maximum likelihood species tree with RAxML under the WAG+GAMMA sequence evolution model with 100 bootstrap replicates.

RESULTS AND DISCUSSION

The final supermatrix included a total of 7,879,645 base pairs, with 48.12% of gap positions or completely undetermined characters. The number of genes per taxon in the final supermatrix is shown in Table 1.

The resulting topology from the maximum likelihood analyses of the supermatrix created by Agalma yields 100 bootstrap support on all nodes; Xylophagidae, Tabanidae, and Pelecorhynchidae, as well as the superfamily Tabanoidea are recovered as monophyletic, as expected, but Rhagionidae does not appear as monophyletic (figure 1); *Symphoromyia sp* is positioned as intermediate between the Xylophagidae and the clade ((*Chrysopilus quadratus* + *Rhigio sp.*) + (Pelecorhynchidae + Tabanidae)), and the clade (*Chrysopilus quadratus* + *Rhagio sp*) is positioned as sister to the Tabanoidea. This result disagrees with those found by Kerr 2010, which, despite the low bootstrap supports, recovered Rhagionidae as monophyletic, and the genera

Symphoromyia and *Chrysopilus* more closely related to each other than any of them to *Rhagio*.

Despite the robust support, the number of taxa in this study is very limited, and a much larger taxon sampling would be necessary in order to truly evaluate the monophyly of Rhagionidae. However, a dataset this size might become intractable with an increased number of taxa, for some steps of the analyses such as the all-by-all blast comparisons are computationally intensive and demand a very long processing time. The size of the concatenated supermatrix of transcriptome data is over 70 times larger than the one generated by Anchored Enrichment and over 1,700 times larger than the one generated by standard molecular techniques, and just as discussed in the previous chapter, one potential pitfall is the possibility of systematic error due to the increased complexity of the dataset and the difficulty to select the correct evolutionary model, which might result strongly supported yet inaccurate phylogenetic reconstruction.

Still, the results of this study are very encouraging in the sense of demonstrating that transcriptome data are able to provide strong phylogenetic signal for deep divergences above the family level in Diptera. Tabanomorpha is a group that diverged around 160Mya, and belongs to Brachycera, a group of flies that underwent rapid diversification and whose relationships remain uncertain (Wiegmann 2011). Future phylogenomic studies may lead to the resolution of this and other regions of rapid radiation of fly lineages, such as lower Diptera (240Mya) and Schizophora (50Mya) (Wiegmann et al 2011).

Overall, Agalma is a very useful, easy to use tool to construct matrices for phylogenomic analyses. The workflow generates a supermatrix in Phylip format, which can easily be used for a primary RAxML analyses through Agalma itself or with any other phylogeny inference software not included in the workflow. It is possible to add new features for specific matrix construction steps, however, intensive bioinformatics skills are required, since the new feature must be developed using BioLite, tested, and once it's working, a request can be sent to the authors in order to incorporate it to the master branch. One particular step that would greatly benefit from alternative methods is the orthology search. Agalma's approach to orthology search consists of identifying homologous sequences by an all-by-all blast comparison, building a gene tree from each set of

homologous sequences, and then identifying the orthologs based on the topologies of gene phylogenies. Orthograph, a method that uses hidden Markov model to develop orthology sets through comparisons with published genome and transcriptome sequences, would be an ideal alternative for a more thorough and model-based orthology detection.

Table 1: Number of genes per taxon in final supermatrix

Species	**Number of genes**
Symphoromyia sp	5225
Pelecorhynchus rubidus	5073
Rhagio sp	4962
Scaptia jacksoniensis	4932
Pelecorhynchus fulvus	4770
Chrysopilus quadratus	4735
Dialysis sp	4616
Haematopota pluvialis	4521
Xylophagus abdominalis	4521
Chrysops flavidus	3871

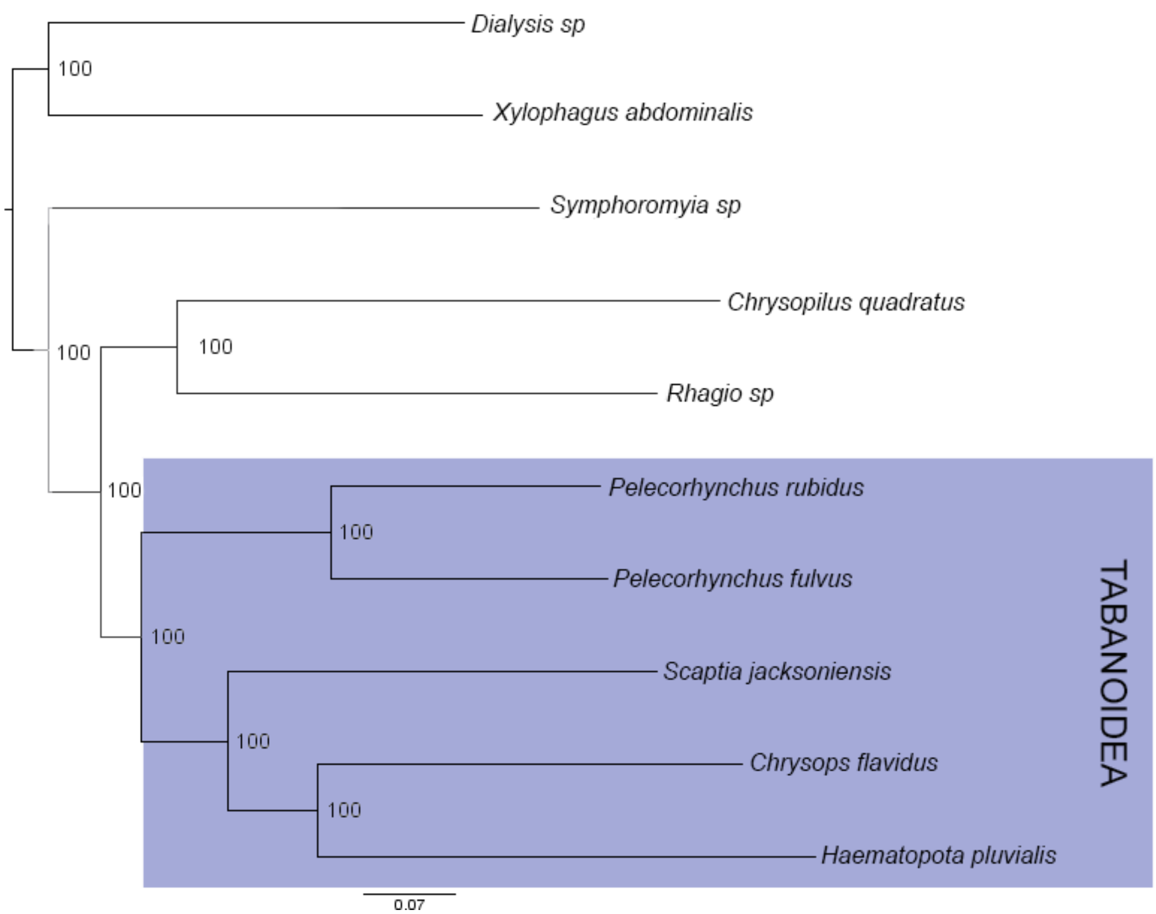

Figure 1: Maximum Likelihood topology based on protein supermatrix, WAG+GAMMA, 100 bootstrap replicates

CONCLUSION

The automated Agalma workflow appears to be a very useful tool for systematists, since all the steps of the analyses are minutely described in the code used to execute the analyses, and its results are easily reproduced because the analyses code won't need to be rewritten. Also, this workflow includes only RAxML for phylogenetic inference, however, the generated supermatrix is in Phylip format and can easily be used with other phylogeny inference software. One possible drawback, however, is the all-by-all blast comparisons, which are computationally intensive and require long processing time, and might become intractable for datasets with a larger number of taxa.

REFERENCES

Bayless, K.M. and Wiegmann, B.M. (2012). *Evolutionary Studies in Tabanomorpha (Diptera) with New Classifications of Xylophagidae and Tabaninae.* Unpublished manuscript. Department of Entomology, North Carolina State University, Raleigh, North Carolina.

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

Dunn, C.W., Howison, M., and Zapata, F. (2013). Agalma: an automated phylogenomics workflow. BMC *Bioinformatics 14*: 330.

Hittinger, C.T., Johnston, M., Tossberg, J.T., Rokas, A. (2010). Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proceedings of the National Academy of Sciences 107*(4): 1476–1481.

Kerr, P.H. (2010). Phylogeny and classification of Rhagionidae, with implications for Tabanomorpha (Diptera: Brachycera). *Zootaxa 2592*: 1–133.

Lemmon, E.M. and Lemmon, A.R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Reviews of Ecology, Evolution and Systematics 44*: 99–121.

Mackerras, I. M. (1954). The classification and distribution of Tabanidae (Diptera) I. General Review. *Australian Journal of Zoology 2*: 431–454.

Ranwez, V., Harispe, S., Delsuc, F., Douzery, E.J.P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS ONE 6*(9): e22594.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics 30*(9):1312-1313.

Stuckenberg, B.R. (2001). Pruning the tree: a critical review of classifications of the Homeodactyla (Diptera, Brachycera), with new perspectives and an alternative classification. *Studia Dipterologica 8*(1), 3–41.

Sukumaran, J. and Holder, M.T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics 26*: 1569-1571.

Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 10*: 57–63.

Wiegmann, B.M., Tsaur, S.C., Webb, D.W., Yeates, D.K., and Cassel, B.K. (2000). Monophyly and relationships of the Tabanomorpha (Diptera: Brachycera) based on 28S ribosomal gene sequences. *Annals of the Entomological Society of America 93*(5), 1031–1038.

Wiegmann, B.M., Trautwein, M., Winkler, I.S., Barr, N.B., Kim, J.W., Lambkin, C., … Yeates, D.K. (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences* March 14, 2011.

APPENDICES

APPENDIX A: Subfamily Chrysopsinae (580 species, 35 genera)

Tribe Bouvieromyini –175 species, 13 genera

1. *Aegophagamyia* Austen, 1912 – 2 subgenera – 37 species (AT)
2. *Eucompsa* Enderlein, 1922 – 2 species (OR)
3. *Gressittia* Philip & Mackerras, 1960 – 12 species (OR; PA; AT)
4. *Lilaea* Walker, 1850 – 5 subgenera – 14 Species (AU)
5. *Merycomyia* Hine, 1912 – 2 species (NA)
6. *Mesomyia* Macquart, 1850 - 26 species (AT; AU)
7. *Nagatomyia* Murdoch & Takashasi, 1961 – 1 species (PA)
8. *Paulianomyia* Oldroyd, 1957 – 1 species (AT)
9. *Phibalomyia* Taylor, 1920– 1 species (AU)
10. *Pseudopangonia* Ricardo, 1915 – 1 species (AU)
11. *Pseudotabanus* Ricardo, 1915 – 2 subgenera – 29 species (NT; AU)
12. *Rhigioglossa* Wiedemann, 1828 – 6 subgenera – 48 species (AT)
13. *Thaumastomyia* Philip & Mackerras, 1960 – 1 species (PA)

Tribe Chrysopsini – 336 species, 8 genera

1. *Chrysops* Meigen, 1803 – 285 species (OR; PA; AT; NT; NA; AU)
2. *Melissomorpha* Ricardo, 1906 – 1 species (OR)
3. *Nemorius* Rondani, 1856 – 7 species (OR; PA)
4. *Nothosilvius* Burger, 1982 – 1 species (OR)
5. *Silviomyza* Philip & Mackerras, 1960 – 1 species (OR)
6. *Silvius* Meigen, 1820 – 7 subgenera – 38 species (OR; PA; NT; NA)
7. *Surcoufia* Krober, 1922 – 1 species (PA)
8. *Neochrysops* Walton, 1918 – 1 species (NA)

Tribe Rhinomyzini –69 species, 14 genera

1. *Alocella* Quentin, 1990 – 1 species (AT)
2. *Betrequia* Oldroyd, 1970 – 1 species (NT)
3. *Gastroxides* Saunders, 1842– 3 species (OR)

4. *Jashinea* Oldroyd, 1970 – 9 species (AT)
5. *Mackerrasia* Travassos Dias, 1956 – 2 species (AT)
6. *Oldroydiella* Travassos Dias, 1955 – 2 species(AT)
7. *Orgizocella* Quentin, 1990 – 2 species (AT)
8. *Orgizomyia* Grunberg, 1906 – 1 species (AT)
9. *Rhinomyza* Wiedemann, 1820 – 4 species (OR)
10. *Seguytabanus* Paulian, 1962 – 1 species (AT)
11. *Sphecodemyia* Austen, 1937 – 5 species (AT)
12. *Tabanocella* Bigot, 1856 – 2 subgenera – 30 species (AT)
13. *Thaumastocera* Grunberg, 1906 – 4 species (AT)
14. *Thriambeutes* Grunberg, 1906 – 4 species (AT)

OR = Oriental; AU = Australia; AT = Afrotropical; PA = Paleartic; NA = Nearctic; NT = Neotropical

APPENDIX B: Taxon Sampling and Gene Coverage (Chapter 1)

Species ID	28S	CO1	CAD1	CAD3	AATS	PGD	16S
Acanthocera sp. 1	YES	NO	NO	YES	NO	NO	NO
Adersia oestroides	YES	YES	YES	NO	YES	NO	YES
Aegophagamyia cincta	NO	YES	NO	YES	NO	NO	YES
Aegophagamyia pulchella	YES	YES	NO	YES	NO	NO	YES
Aegophagamyia sp. 1	YES	YES	YES	NO	YES	NO	YES
Aegophagamyia variegata	YES	YES	NO	NO	YES	NO	NO
Agelanius lanei	YES	YES	YES	NO	YES	NO	NO
Apatolestes rugosus	YES	YES	YES	NO	YES	YES	NO
Asaphomyia floridensis	YES	NO	YES	NO	YES	NO	NO
Bolbodimyia brunneipennis	NO	YES	YES	YES	NO	YES	NO
Chasmia sp. 1	YES	YES	YES	NO	YES	YES	NO
Chlorotabanus mexicanus	NO	YES	YES	YES	YES	YES	NO
Chrysops caecutiens	YES	YES	NO	NO	YES	NO	YES
Chrysops callidus	YES	YES	NO	NO	NO	NO	NO
Chrysops coquilletti	YES	YES	YES	NO	YES	NO	NO
Chrysops dispar	YES	YES	NO	NO	YES	NO	YES
Chrysops flavidus	YES	YES	NO	NO	YES	NO	NO
Chrysops funebris	YES	YES	NO	NO	NO	NO	YES
Chrysops indus	YES	YES	NO	NO	NO	NO	NO
Chrysops longicornis	YES	YES	NO	NO	NO	NO	NO
Chrysops macquarti	NO	YES	NO	NO	NO	NO	NO
Chrysops madagascariensis	YES	YES	YES	NO	NO	NO	NO
Chrysops moechus	NO	YES	NO	NO	NO	NO	NO
Chrysops noctifera	NO	YES	NO	NO	NO	NO	YES
Chrysops sp. 20	YES	YES	NO	NO	NO	NO	NO
Chrysops sp. M02	YES	YES	NO	NO	YES	NO	YES

Species ID	**28S**	**CO1**	**CAD1**	**CAD3**	**AATS**	**PGD**	**16S**
Chrysops sp. M03	NO	YES	YES	NO	YES	NO	YES
Chrysops sp. M04	NO	YES	NO	NO	YES	NO	YES
Chrysops sp. M05	YES	YES	NO	NO	YES	NO	YES
Chrysops sp. M06	YES	NO	NO	NO	NO	NO	YES
Chrysops sp. M07	YES	YES	NO	NO	NO	NO	NO
Chrysops sp. M09	YES	YES	NO	NO	NO	NO	YES
Chrysops sp. M10	YES	YES	NO	NO	NO	NO	YES
Chrysops sp. M11	YES	YES	NO	NO	NO	NO	YES
Chrysops sp. M13	YES	YES	NO	NO	NO	NO	NO
Chrysops sp. M14	YES	YES	NO	NO	NO	NO	YES
Chrysops sp. M15	YES	NO	NO	NO	NO	NO	YES
Chrysops upsilon	YES	YES	NO	NO	NO	NO	NO
Chrysops virgulatus	YES	YES	NO	NO	NO	NO	YES
Chrysops vittatus	YES	NO	NO	NO	YES	NO	NO
Coracella sp. 1	YES	YES	NO	NO	YES	NO	YES
Dasybasis sp. 12	YES	YES	YES	YES	YES	YES	NO
Dasychela peruviana	YES	YES	YES	YES	YES	YES	NO
Dasyrhamphis umbrinus	YES	YES	YES	YES	YES	YES	NO
Diachlorus bimaculatus	YES	YES	YES	YES	YES	YES	NO
Dicladocera macula	NO	YES	YES	YES	YES	NO	NO
Ectenopsis sp. 2	YES	YES	NO	NO	YES	NO	NO
Ectenopsis vulpecula	YES	YES	NO	NO	YES	YES	NO
Esenbeckia lugubris	YES	YES	NO	YES	YES	NO	NO
Esenbeckia prasiniventris	YES	YES	YES	YES	YES	NO	NO
Goniops chrysocoma	NO	YES	YES	YES	YES	YES	NO
Leucotabanus albovarius	NO	YES	YES	YES	YES	YES	NO
Lilaea fuliginosa	YES	YES	NO	NO	YES	NO	YES
Lilaea sp. 3	YES	YES	NO	YES	YES	YES	YES

Species ID	28S	CO1	CAD1	CAD3	AATS	PGD	16S
Limata tenuicornis	YES	YES	NO	YES	YES	YES	NO
Mycteromyia sp. 1	YES	YES	YES	NO	YES	NO	NO
Orgizomyia zigzag	YES	YES	YES	YES	NO	YES	YES
Pegasomyia abaureus	YES	YES	YES	NO	NO	NO	NO
Philoliche alternans	YES	YES	YES	NO	YES	NO	NO
Philoliche basalis	YES	YES	YES	NO	YES	NO	NO
Pityocera cervus	YES	YES	YES	NO	YES	NO	NO
Poeciloderas quadripunctatus	YES	YES	NO	YES	YES	YES	NO
Pseudotabanus sp. 1	YES	YES	NO	YES	YES	YES	NO
Pseudotabanus sp. 2	NO	YES	NO	NO	NO	NO	YES
Rhigioglossa edentula	YES	YES	YES	YES	YES	NO	NO
Rhigioglossa obtundata	YES	YES	NO	NO	NO	NO	YES
Rhigioglossa sp. 2	YES	YES	NO	NO	NO	NO	YES
Scione maculipennis	YES	YES	YES	NO	YES	NO	NO
Silvius gigantulus	YES	YES	YES	NO	YES	NO	NO
Silvius quadrivittatus	YES	YES	NO	NO	NO	NO	YES
Silvius variegatus	YES	YES	YES	YES	YES	YES	YES
Stenotabanus sp. 5	YES	YES	YES	YES	YES	YES	NO
Stonemyia rasa	YES	YES	YES	NO	YES	NO	NO
Tabanocella denticornis	YES	YES	YES	NO	YES	NO	NO
Tabanocella longirostris	YES	YES	NO	YES	YES	YES	YES
Tabanocella sp. 3	YES	YES	NO	YES	YES	YES	YES
Tabanus sudeticus	YES	YES	YES	NO	YES	YES	NO
Thaumastocera akwa	YES	YES	YES	YES	YES	NO	YES
Triclista guttata	YES	YES	YES	NO	YES	NO	NO

APPENDIX C: Taxa Provenance (Chapters 1 and 2)

Species ID	Locality Verbatim
Acanthocera sp. 1	Ecuador: Yasuni: T. Erwin
Adersia oestroides	Kenya, Coast Province Msambweni, on beach 4°28.41'S. 39°29.44'E 3.I.2000, sweep net R. Copeland
Aegophagamyia cincta	Madagascar: Diego Prov, Sakaramy, 7km N Joffre Ville; hand net in humid tropical forest; 7.i.2007; M.E. Irwin, F.D. Parker; 425m; 12°26.77'S, 49°14.10'E
Aegophagamyia pulchella	Madagascar: Majunga Prov, 20km NW Port. Berger, Ambovomamy Belambo; hand netted in secondary growth on white sand; 4.1.2007; M.E. Irwin, F.D. Parker, R. Harin Hala; 36m; 15°27.07'S 47°36.80'E
Aegophagamyia sp. 1	Madagascar: Antsiranana Province, Sakarany, 7km N. Jaffreville hand netted in tropical deciduous forest; M.E.Irwin, F.D. Parker, R.Harin'Hala 2-7.I.2007; 360m; 12°20'S 49°15'E
Aegophagamyia variegata	Madagascar: Antsiranana Province, Sakarany, 7km N. Jaffreville hand netted in tropical deciduous forest; M.E.Irwin, F.D. Parker, R.Harin'Hala 2-7.I.2007; 360m; 12°20'S 49°15'E
Agelanius lanei	Chile: PNAA: Chaicra Trap 3 I.2007 Admin [B. Brown col]
Apatolestes rugosus	USA: CA: Mendocino Co.; Hopland REC, Lookout Point hilltop; 898m; 39°01'36.0"N 123°03'35.6"W; 30.v.2009; J. F. Gibson & O. Lonsdale
Asaphomyia floridensis	USA: FL: M. Deyrup

Bolbodimyia brunneipennis	French Guiana: Regina; Kaw Mtns, Point Road 40 (PR40) (Rainforest #1 Plantation #3) ca. 300m, 3-27.iii.2006, MT leg. Keijo Sarv, sample ID FR-GY/2006/108-(Tabanidae) sorted by Marc Pollet
Chasmia sp. 1	Papua New Guinea: Chimbu Province; Wara Sera Research Station Crater Mountain Research Area 8-13.vii.2001 [Bradler, Jarvis, Svenson cols]
Chlorotabanus mexicanus	Costa Rica: Heredia Ur. Puerto Viejo, La Selva Biol. Sta. 179'; 10°25'N 84°00'W; 23.ii-2.iii.2004 K.B. Miller
Chrysops caecutiens	Serbia: Frusca Gora National Park; Meadow above Conference Centre; 20.vi.2009; 45.160247°N 19.647778°E; 276m; J. Skevington
Chrysops callidus	USA: NC: Wake Co: Raleigh; Schenck Forest day sweep human bait 4.v.2009 K. M. Bayless
Chrysops coquilletti	CA:USA [E. Sarnat col]
Chrysops dispar	Thailand: Chaiyaphum; Pa Hin Ngam NP Mixed deciduous forest; 15°34.686N 101°26.082E 419m Malaise trap 10-16.xi.2006; Katae Sa-nog & Buakaw Adnafai leg. T1033
Chrysops flavidus	USA: NC: Wake Co: Raleigh; Schenck Forest day sweep 19.vi.2008 KM Bayless
Chrysops funebris	[Kenya]: Kakamega forest; Malaise trap; 17-31 July 2006; 95%
Chrysops indus	Canada:ON: Burnt Lands Provincial Park Reserve near Almonte; 45.248998°N 76°145996°W Malaise Trap; 4-11.vi.2009; M. M. Locke
Chrysops longicornis	Kenya: Njuki-ini Forest; 10-24.xi.2008

Chrysops macquarti	USA: NC: Wake Co: Raleigh; Schenck Forest; Malaise Trap Stream Bed; 29.v-7.vi.2009 K. M. Bayless
Chrysops madagascariensis	Madagascar: Majunga Prov. 20km NW Port-Berger, Ambovomamy Belambo, hand-netted in secondary growth on white sand; 9.I.2007; M.E.Irwin, F.D. Parker, R.Harin'Hala; 36m;
Chrysops moechus	USA: NC: Durham Co: Hill Demonstration Forest: MT mixed forest 500m from 36.201°N 78.887°W A. R. Deans 3.vi.2009
Chrysops noctifera	USA: WA: Wawawa County Park on Snake River; marshy, weedy area; 46°38.187'N 114°22.479'W 9.vi.2009 I. S. Winkler; E09-61
Chrysops sp. 20	Peru:Rio Tambopata Explorers Inn- Rio Tower 12°50.208S 069°17.603W 6-13.xii.2003 Svenson, Miller, Osborne, Ogden
Chrysops sp. M02	BR: AM: Presidente Figueiredo, AM 240, Km 24, 30.ix-3.x.2010
Chrysops sp. M03	Mongolia: Selenge Aimag: Zuunburen Soum: small channel of Selenge River, 7.9 km SW of Zuunburen; N 50.10277° E 105.78217° 627m elev.; 6 July 2005; C.R. Nelson #8182 & SRP crew, SRP #0507501
Chrysops sp. M04	Mongolia: Bulgan Aimag: Teshig Soum, West Branch of Taria Khtain Go; Stream; N 49.70184° E 130.80040°; elev 1436m; 9 July 2005; C. R. Nelson #8195 & SRP #0507901
Chrysops sp. M05	Mongolia: Housgol Aimag; Ulaan Uul Soum, Soyo Brigade, springs of Shishged Gol; 36 km NNE of Ulaan Uul town; N 50.99171° E 99.35400°; elev 1584m; 30

	June 2006; C. R. Nelson #8473 & SRP #2006063002
Chrysops sp. M06	Bolivia: Buena Vista; 23.II.99; Malaise; F.D. Parker
Chrysops sp. M07	Panama: CI Barro Colorado Isl.; 9°9'N 79°51'W; 15-22.ix.1995; J. Pickering
Chrysops sp. M09	Costa Rica: Alajuela: Penas Blancas, San Ramón; 450m; 31.iii.2010; Malaise; W. Porras; N 843705
Chrysops sp. M10	Bolivia: Buena Vista; 21.II.99; Malaise; F.D. Parker
Chrysops sp. M11	Peru: Amazonia Lodge; Malaise trap; 12°52'16"S 71°22'11"W; 23-24.x.2006; a 465m; J. Skevington
Chrysops sp. M13	Chile: Limari prov. Fray Jorge Natl Park Queb. Las vacas 2km NW Admin. 1-3.Oct.97. Pantrap; Irwin, Schlinger; 170m; 30.6710°S 71.6121°W
Chrysops sp. M14	Costa Rica, Puntarenas Prov., Coto Brus Cant., Las Cruces Reserve, 2mi S of San Vito, 8.785N, 82.959W, 4.viii.2010, B. Wiegmann et al.
Chrysops sp. M15	Venezuela: Lara: Yacambu National Park, Headquarters; 5-8.ix.2008; Malaise trap 9°42'25"N, 69°34'38"W; J. Skevington; 1549m
Chrysops upsilon	Maryland: Anne Arundel Co; Jug Bay Wetlands Sanctuary Glendening Preserve, Sand Barrens ("The Desert") 38°48.402'N 76°42.097'W 9-25.vii.2007 Gates, Buffington, Kula
Chrysops virgulatus	Mexico: Sonora: Rancho el Cajon; 40km E of Alamos: Malaise on Sand, bench of Rio Cuchujaqui; Taxodium riparian tropical deciduous forest 1-11.x.2006; M. E. Irwin; 420m 27°03.00'N 108°43.91'W

Chrysops vittatus	Illinois [Wiegmann col]
Coracella sp. 1	Chile, Quillota Province; Palma de Ocoa; Parque Nacional Campanas malaise in hillside draw; 215 m; 2/10-I-2000 M.E. Irwin, E. I. Schlinger 32.9324°S 71.0781°W
Dasybasis sp. 12	Peru: Between Cusco & Peema on bus; 20-08-10 B. Lessard
Dasychela peruviana	Peru: Cock-of-the-Rock Lodge; Malaise trap; 13°03'21"S, 71°32'46"W; 18-20.x.2006; ~1380m J. Skevington
Dasyrhamphis umbrinus	Greece: Lesvos Island: nr. Mystegna 319.v.2009 leg G. Ståhls
Diachlorus bimaculatus	Peru: Rio Tambopata Explorers Inn 12°50.208'S 069°17.603'W; 2-13.viii.2005 Whiting et al. Canopy Light Trap
Dicladocera macula	[Ecuador] Pichincha: Yanacocha; paramo; parking lot sweep off cars; 00°06.693'S 78°35.082'W;3539m;3.viii.2009; KMB09122; D.E.Ramirez, K.M.Bayless, R.Cardenas
Ectenopsis sp. 2	Australia: Northern Territory Mary River- roadhouse camp; South edge of Kakadu NP 13°36'47.1" 132°13'41.3"E 27.ix.2002 Whiting, Ogden, Svenson
Ectenopsis vulpecula	Australia: Old Bribie Island; S27.053' W153.180' 17.x.2007 SL Winterton; Heathland
Esenbeckia lugubris	Bolivia: Santa Cruz Co.; San Jose; MVL; 17°44.903'S 061°04.202'W 13.I.2004 Svenson, Cameron, Bybee
Esenbeckia prasiniventris	Costa Rica: Caribe: Rio francesca Malaise trap; 11-17.vi.2007 ARDeans

Goniops chrysocoma	USA: Maryland: Little Orleans: 2007
Leucotabanus albovarius	[Ecuador] Alinahui: 5.iii.2006 R. Cardenas
Lilaea fuliginosa	AUS: NSW; Tallaganda Natl Pk; 7.8 km E Hoskinstown; Malaise trap 9-24 Jan 2006 1142m S. Fayed, C. Manchester, S. Alasya, ANIC Bulk sample No 2824 35°24'53"S 149°32'07"E GPS
Lilaea sp. 3	Australia: Northern Territory Kakadu National Park-Gabiru 12°50'37.2"S 132°51'26.9"E 28.ix.2002 Whiting, Ogden, Svenson
Limata tenuicornis	RSA: Western Cape, vegetated dunes+margin S Papendorp, 31°43'01"S 018°12'22"E 9m 29.ix.2009 T. Dikow & J. Londt
Merycomyia whitneyi	Log in H_2O; 4.3.52; Ben Ranch Johnson Flo.; Pupated 4.27.52; Emerged 5.15.52; Coll. C. M. Jones
Mycteromyia sp. 1	Chile [Wiegmann col]
Neochrysops globosus	Newark, Del. 3 July 1967; E. P. Catts; at edge of woods on aster
Orgizomyia zigzag	Madagascar: Antsiranana Province, Sakarany, 7km N. Jaffreville hand netted in tropical deciduous forest; M.E.Irwin, F.D. Parker, R.Harin'Hala 2-7.I.2007; 360m; 12°20;S 49°15;E
Pegasomyia abaureus	CA:USA [P.S. Cranston col]
Philoliche alternans	Westcoast Fossil Park RSA [M. Buffington col]
Philoliche basalis	Neuwoutville, RSA [S. Van Noort col]
Pityocera cervus	Peru:Rio Tambopata Explorers Inn- Rio Tower 12°50.208S 069°17.603W 6-13.xii.2003 Svenson, Cameron, Bybee

Poeciloderas quadripunctatus	[Ecuador] Alinahui: 8.iii.2006 R. Cardenas
Pseudotabanus sp. 1	Australia: SE Queensland; Brisbane Forest pk, Scrub Rd; S27.427 E152.841 13.xii.2007-17.I.2008; Malaise over shallow creek in rainforest; S.L. Winterton & J.S. Bartlett
Pseudotabanus sp. 2	Australia: B. D. Lessard
Rhigioglossa edentula	South Africa: WCP: S. I. Morita
Rhigioglossa obtundata	South Africa, Western cape, 4 Km S. Prince Albert; Swartberg Mtusi; malaise in wash with spring; 700m; 2-23.x.2004; M.E. Irwin, F.D. Parker, M. Hauser; 33°16'00"S, 22°02'47"E
Rhigioglossa sp. 2	RSA: W Cape: 7 km E Ladismith; MT in sandy wash 1-23.x.04 520m ME Irwin, M Hauser, FD Parker 33°31'00"S 21°19'51"E
Scione maculipennis	CR: Cacao La Cim. Malaise A.R. Deans 19-iv-1999
Silvius gigantulus	CA:USA [S. I. Morita col]
Silvius quadrivittatus	USA: Arizona: Maricopa Co: Gila River nr. Gillespie Dam, 8km S. of Arlington, malaise in Tamarisk thicket; 250m; 28.iv-4.v.2007 ME Irwin 33°13.63N 112°46.15W
Silvius variegatus	Spain: Valencia: El Honda Reserve; 3m 38°12'13''N 00°45'24''W 29.v.2008 J. & R. Skevington
Stenotabanus sp. 5	[Ecuador] Pichincha: Nambillo: Mindo: El Monte Lodge: cloud forest malaise trap; 00°04.169'S 78°45.741'W 1307m; 29.vii.2009 K.M.Bayless, D.E.Ramirez, R.Cardenas
Stonemyia rasa	USA: NC: Wake Co: Raleigh; Schenck Forest MT 26.v-3.vi.2008 I. Winkler

Tabanocella denticornis	KZN: South Africa [S. I. Morita col]
Tabanocella longirostris	[Madagascar] Ambovomamy Belambo, hand-netted in secondary growth on white sand; 9.I.2007; M.E.Irwin, F.D. Parker, R.Harin'Hala; 36m; 15°27.07'S 47°36.80'E
Tabanocella sp. 3	Zambia: Copperbelt Province; Kitwe District; Kumasamba Lodge Near Kafue River, Abt. 5 km S Kitwe; 12°54'20"S 28°14'22.3"E 18-30.iv.2006 G. Svenson
Tabanus sudeticus	Greece: Kerkini; Ramna Site; M.T. N41°17'42.5 E023°11'33.1 25.v-1.vi.2008 T. Zeegers
Thaumastocera akwa	Nigeria; IITA Forest MT iii.06
Triclista guttata	AUSTRALIA: NSW, Nelligen, 35°38'52"S 150°08'29"E, 11 Feb 2008 [D. Yeates col]

APPENDIX D: Primers (Chapter 1)

Primer	Sequence
CO1	
Ty-J-1460F	TACAATTTATCGCCTAAACTTCAGCC
C1-N-2329R	ACTGTAAATATATGATGAGCTCA
ChryCO1f	TTTGAYCGWATACCHYTAT
ChryCO1r	ATARDGGTATSCGRTCAAA
HCO 2198R	TAAACTTCAGGGTGACCAAAAAATCA
LCO 1490F	GGTCAACAAATCATAAAGATATTGG
28S	
28h	GGTTTCGCTGGATAGTAG
28i	GGGTCTTCTTTCCCCGCT
28K	GAAGAGCCGACATCGAAG
28zb	GCAAAGGATAAGCTTCAGTGGAT
28zc	TGGATCGCAGTATGGCAGCT
rc28d	CCGCAGCTGGTCTCCAAG
rc28F	GTGATTTCTGCCCAGTGCTCTG
rc28H	CTACTATCCAGCGAAACC
rc28Q	GGACATTGCCAGGTAGGGAGTT
CAD1	
CAD 54F	GTNGTNTTYCARACNGGNATGGT
166'F tab	CGCATAATTCATCA
384R tab	DCCTTCRTTDGSNTKRTCRTTNABRTTHG
ChryCAD 350R	GTRTCVTGATGAATRGAVGG
ChryCAD 364R	TCYACNGCAAAHCCRTGRTTYTG
ChryCAD1F1	CCDARRCATTTYGARTGG

ChryCAD1F2	TAYAGRTCRCAAATHYTWGT
CAD3	
CAD 581*F	GGWGGWCAAACWGCWYTMAAYTGYGG
CAD 833 R	GCCATNACYTCSCCNACACTYTTCAT
ChryCAD 603F	CTTGGHACKCCRATYAAATCAAT
ChryCAD 606F	CCRATYAAATCAATYATAGA
ChryCAD 800R	ACKACRCARTAATCBARRCTDGGTTC
ChryCAD 804R	GGSATYTTNACKACRCARTAA
ChryCAD 812R	GTRAAYTTNGCYAARTCCCARCG
M13CAD 800R dia	CAGGAAACAGCTATGACACSACACARTAATCSARRCTVGGTTC
AATS	
M13A1X244R tab	CAGGAAACAGCTATGACCATKCCGCARTCRATGTGTTT
M13A1X242R tab	CAGGAAACAGCTATGACGCARTCRATGTGTTTYTTBGG
M13A1X54F tab	TGTAAAACGACGGCCAGTAAYAGYGAYATGGCRAARTGG
ChryAATS1F1	CMAAYACNCARAARTGYAT
ChryAATS1F2	ACYAGYGAYATGGCRAARTGG
ChryAATS1R1	TTCCADAYTTCSAGTACATC
ChryAATS1R2	ACATCSGGATCRTCCATRTT
PGD	
4R-31R dia	CCCCACTTAACYGCRTGACTAACAAC
ChryPGD F1	ATAARTGGAATAWRGARGA
ChryPGD F2	ATATWYTVAAKTAYAARGAT
ChryPGD R1	CCTTCGAATRATRCARCCTCCAC
ChryPGD R2	GATTCTTGACCTTTYTCAAT
PGD 2+21F dia	GCTGYYGARTTYGATAAATGGAA

M13 PGD 4R	CAGGAAACAGCTATGACCCNGTCCARTTNGTRTG
16S	
LR-J-12887	CCGGTTTGAACTCAGATCATGT
LR-N-13398	CGCCTGTTTAACAAAAACAT
Chry16Sf	AATTTTATTRTCACCCCAA
Chry16Sr	GACGAGAAGACCCTATAGAT