Building the Dipteran Tree: Co-operative Research in Phylogenetics and Bioinformatics of True Flies (Insecta: Diptera)

===

I. Response to Previous Reviewers

This proposal is a resubmission. Below is a list of major modifications and responses to issues raised by the panel and reviewers:

Project is too ambitious and PIs are overcommitted. We have recruited one additional laboratory with access to high throughput molecular techniques for the mt genome data collection which makes the determination of complete mt genomes more straightforward and routine. The workplan for the morphological data collection is now described in a separate section and several projects of PIs will end within the first year of the current request. We are optimistic that we will be able to meet the project goals given that the PIs have within the past five years published cladistic analyses covering more than 260 species in 15families and that we have for the current request recruited the help of more than 15 collaborators, students, and postdocs.

Problematic taxon sampling. We will use the current family classification as sampling scheme for the second tier analysis, because the vast majority of Diptera families have been documented to be monophyletic (see ref. in text); and, for those of uncertain monophyly, we propose the inclusion of multiple exemplars. As suggested by the panel, we have increased the taxon sample for the Nematocera and will give high priority to "problematic" families. Furthermore, our analysis will be based on exemplars and thus does not assume the monophyly of any higher-level taxon.

Unclear roles of project participants. More details are now provided (see Letters of support and Management plan). Some collaborators mostly assist with specimen acquisition, others are actively engaged in building the data sets. We purposefully built an extensive international network of collaborators from all continents in order to assure that we can deliver good taxon sampling, avoid unnecessary travel costs, and circumvent problems with collecting permits.

Not enough preliminary data. We have included significant new preliminary data, including: (1) a preliminary analysis of 5 genes for 37 of the 42 first tier taxa; (2) phylogenetic analysis of 10 complete mitochondrial genomes; and (3) a Diptera supertree (see figures 1- 3).

Not enough methodological innovation and expertise. Our proposal explicitly falls under the "taxon focus" category in the ATOL program description. We believe that given the sampling intensity demanded by a megadiverse insect order like the Diptera, only a taxon-focused project can be successful. The project nevertheless addresses important theoretical issues such as the relative merit of intensive taxon- versus character- sampling and the feasibility of supertree reconstruction. The methodological expertise of the PIs is documented by numerous publications in leading journals (see CVs) and the organization of symposia at international conferences. We have furthermore recently reconstructed a first dipteran supertree (see below).

II. Results from Prior NSF Support (last 5 years)

Co-PIs Wiegmann and Yeates: DEB-PEET 9521925, 9977958. Towards a world monograph of the Therevidae (Diptera). M.E. Irwin, B.M. Wiegmann, and D. Yeates, Co-PIs. 9/1/95-8/31/00; $725,000 ($145,000 to BMW). Renewed 9/1/2000-8/31/2005 ($150,000 to BMW; $180,000 to DKY). This project combines revisionary taxonomic work with morphological and molecular phylogenetic data collection to assess relationships within the family, and among its closest relatives. Six project graduate students have completed their degrees and one additional graduate student is currently supported. **Publications: (partial project listing):**

Hill, H.N. (2003). Investigation of the phylogenetic utility of two nuclear genes, Opsin and CAD, within the stiletto flies (Diptera: Therevidae). M.S. Thesis. (**B. M. Wiegmann**, Advisor).

Yeates, D.K., Irwin, M.E., B.M. Wiegmann. (2003) Ocoidae, a new family of asiloid flies (Diptera: Brachycera: Asiloidea), based on *Ocoa chilensis gen. and sp. nov.* from Chile, South America. *Syst. Entomol.*, in press.

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

Winterton, S.L., Yang, L., Wiegmann, B.M., Yeates, D.K. (2001) Phylogenetic revision of the Agapophytinae subfam. n. (Diptera: Therevidae) based on molecular and morphological evidence. *Syst. Entomol.* 26: 173-211.

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

Yang, L., Wiegmann, B.M., Yeates, D.K. Irwin, M.E. (2000) Higher-level phylogenetic reconstruction of the Therevidae based on nucleotide sequence data. *Mol. Phyl. & Evol.* 15: 440-451.

Wiegmann, B.M., Tsaur, S.C., Webb, D.W., Yeates, D.K., Cassel, B.K. (2000) Monophyly and relationships of the Tabanomorpha (Diptera: Brachycera) based on 28S ribosomal gene sequences. *Ann. Entomol. Soc. Am.* 93: 1031-1038.

Yeates, D.K., Wiegmann, B.M. (1999) Congruence and controversy: toward a higher-level Phylogeny of the Diptera. *Annu. Rev. Entomol.* 44: 397-428.

Co-PI Wiegmann: DEB 0089745, Molecular phylogenetics and a time scale for diversification of the higher flies (Diptera: Eremoneura). B.M. Wiegmann, J.L. Thorne Co-PIs. 4/15/01-04/14/04; $255,000. This project focuses on the phylogeny of the higher dipteran group, Eremoneura, and applies molecular, morphological and paleontological evidence in divergence time estimation using Bayesian methods.

Publications supported by the proposal:

Wiegmann, B.M., Yeates, D.K., Thorne, J.L., Kishino, H. (2003) Time Flies: A new molecular time-scale for fly evolution without a clock. *Syst. Biol.* in press.

Moulton, J.K., Wiegmann, B.M. Evolution and phylogenetic utility of CAD (Rudimentary) among Mesozoic-aged eremoneuran Diptera. *Syst. Biol.* submitted.

Collins, K.P. and Wiegmann, B.M. (2002) Phylogenetic relationships of the Eremoneura (Diptera) based on 28S rDNA and EF-1α sequences. *Ins. Syst. & Evol.* 33:421-444.

Collins, K.P. and Wiegmann, B.M. (2002) Phylogenetic relationships of the lower Cyclorrhapha (Diptera:) based on 28S rDNA sequences. *Ins. Syst. & Evol.*, 33:445-456.

Co-PI Courtney: DEB-9407153 (DEB-9796275), Systematics, cytogenetics and life histories of Nearctic *Blepharicera* (Diptera: Blephariceridae). G.W. Courtney (PI). 02/01/95-07/31/99. $90,000. This project included revisionary systematics, chromosomal and phenological studies, and cladistic analysis of Nearctic species of *Blepharicera*. The project supported two graduate students (Clemson University, Iowa State University), and helped establishing a blephariceridwebsite (http://www.ent.iastate.edu/dept/research/systematics/bleph/). **Publications**:

Alverson, A.J. & G.W. Courtney. (2002) Temporal patterns of diatom ingestion by larval net-winged midges (Diptera: Blephariceridae: *Blepharicera*). *Freshwater Biology* 47: 2087-2097

Alverson, A.J., G.W. Courtney, & M.R. Luttenton. (2001) Niche overlap of sympatric *Blepharicera* larvae (Diptera: Blephariceridae) from the southern Appalachian Mountains. *J. North Am. Benth. Soc.* 20: 564-581.

Courtney, G.W. (1998) A method for rearing pupae of net-winged midges (Diptera: Blephariceridae) and other torrenticolous flies. *Proc. Entomol. Soc., Wash.* 100: 742-745.

Courtney, G.W. (2000) Revision of the net-winged midges of the genus Blepharicera Macquart (Diptera: Blephariceridae) of eastern North America. *Mem. Entomol. Soc. Wash.* 23: 1-99.

Courtney, G.W. (2000) A.1. Family Blephariceridae. pp. 7-30 in L. Papp & B. Darvas (editors). *Contributions to a Manual of Palaearctic Diptera*. *Appendix.* Science Herald, Budapest.

Courtney, G.W. & R.M. Duffield. (2000) Net-winged midges (Diptera: Blephariceridae): a food resource for Brook Trout in montane streams. *Pan-Pacific Entomol.* 76: 87-94.

Courtney, G.W., B. J. Sinclair & R. Meier. (2000) 1.4. Morphology and terminology of Diptera larvae. pp. 85-161 in L. Papp & B. Darvas (editors). *Contributions to a Manual of Palaearctic Diptera. Volume 1*. Science Herald, Budapest.

Co-PI Courtney: DEB-0103144, Biodiversity of lotic aquatic insects of northern Thailand. G.W. Courtney (PI). 09/01/01-08/31/04. $299,957. This study supports six graduate students and includes inventories of all lotic insects, revisionary systematics of selected taxa, the collection of ecological and distributional data for resident species, the construction of web-based and traditional resources for faunal, distributional, and ecological data. Thus far, research activities have focused on field work and on the processing of collected material. Although sorting and identification continues, the project already has led to the discovery of many new species and the compilation of an extensive reference collection.

Co-PI Friedrich: DBI-0070099, A Laser Scanning Confocal Microscope for Biological Imaging. M. Friedrich, J.E. Lilien, L.A. Elferink, M. VanBerkum, A. Popadic (Co-PIs). 1/6/00-7/31/01; $200,000.

Co-PI Friedrich: DBI-0091926, Functional Analysis of Eye Development in the Grasshopper. 1/1/01-12/31/03; $308,140 Comparative and functional analysis of the signaling genes *decapentaplegic*, *wingless* and *hedgehog* during embryonic eye development in the grasshopper *Schistocerca americana*. **Publications:**

Friedrich M. From lobes to discs: evolution of insect eye development. *Am. Zool.,* submitted.

III. Introduction

It is estimated that currently 1.7 million species are known to science. Of these, 150,000 belong to the Diptera (true flies), one of the four megadiverse orders of insects. Resolving the phylogenetic relationships of Diptera is thus one of the foremost tasks in completing the tree-of-life. We propose a collaborative effort among a large, international team with morphological and molecular expertise to reconstruct the Diptera tree. Despite 40 years of phylogenetic research a well-supported tree for the entire order has yet to be proposed. This situation seriously limits comparative work, although the large number of important dipteran model species has arguably generated more biological data for this order than for any other insect lineage.

We propose a three-tier approach. (1) A first large-scale analysis will be based on a small taxon sample of 42 species representing all major dipteran lineages, a first comprehensive morphological data set across the entire order, and a large amount of DNA data. This analysis is modeled on the recently published, successful, large genomic sample for mammal phylogeny (Murphy et al. 2001). (2) For the second-tier analysis, we will study at least one representative for each of the 160 dipteran families, which will be scored for both molecular and morphological data. This second tier analysis is analogous to the large, successful, collaborative efforts on seed plant phylogeny (Chase et al. 1993; Rice et al. 1997; "Deep Green", Palevitz, 2001). (3) The third tier will rely on the dipteran backbone trees reconstructed in (1) and (2) for constructing a supertree for 1500-2000 species of Diptera

The worldwide web contains thousands of webpages with important information on Diptera. We will build the 'Diptera Biodiversity Web' which will give access to a dramatically expanded Diptera 'Tree-of-Life,' links to important electronic catalogues and collection homepages, databases for Diptera species in culture, and educational information on flies. Furthermore, an existing specimen database program (MANDALA) will link DNA sequences with digital images and supplementary information on the specimens from which the DNA was extracted.

The results of the proposed research will have a tremendous impact on many areas of biology. (1) Comparative biology. The dipteran tree will connect existing information on model species (*Anopheles, Drosophila, Glossina* etc.) and promote comparative research on a large variety of phenomena such as behavior, neuroanatomy, and wing evolution. The tree will be particularly important for evolutionary developmental biology ('evo-devo') in providing context for key developmental events (Schmidt-Ott et al. 2002; Stauber et. al.2002; Hurley et al. 2001; Kopp et al. 2000; Rohr et al. 1999), essential for addressing longstanding homology questions in Diptera morphology (e.g., Sinclair 2000; Sinclair et al. 1994; Courtney et al., 2000), and for understanding the evolution of insecticide resistance (Newcomb et al., 1997; Oakeshott et al., 2003; Ranson et al., 2002; see letter of support by Oakeshott). (2) Genomics. The project will generate complete mitochondrial sequences for the 42 species of the first-tier, thus allowing for the microevolutionary studies of mt genome evolution in *Drosophila* to be extended to deeper phylogenetic levels (Ballard 2000). Furthermore, the estimated 40kb of nuclear DNA sequences will make Diptera a model invertebrate taxon for comparative genomics (Zdobnov et al. 2002; Kaufman et al. 2002; Holt et al. 2002). However, this work will be dependent on the well-founded phylogenetic framework that we propose to generate. (3) Phylogenetic analysis. We will explore two approaches to reconstructing the phylogenies of megadiverse taxa. First, a very large amount of data for a small comprehensive taxon sample (first-tier) will be collected and compared with regard to tree stability and cost-efficiency with a second approach relying on a moderate amount of data for a dramatically improved taxon sample (second-tier). In the third tier, we will furthermore test different techniques for constructing 'supertrees.' (4) Fossil Record and Time of Divergence. Due to excellent information on the fossil record of Diptera, a robust phylogenetic hypothesis will propel the order into a model for integrating information from both fossil and extant diversity.

The project also transcends several academic barriers and reaches out to a diverse student body. (1) Synthetic morphological research across the order will be promoted by generating character definitions that are applicable across the entire order. (2) The project integrates the international Diptera community by involving collaborators from all continents. (3) The project promotes synergy between morphological and molecular systematics by offering student internships in molecular systematics for students with morphological thesis projects.

III.1 Specific Objectives

1) Generate a robust family-level phylogenetic hypothesis for Diptera based on a character matrix consisting of morphological characters and molecular sequences
2) Build a Diptera supertree for more than 1500 species based on new and published trees
3) Compare the phylogenetic utility of the mitochondrial and nuclear genomes at multiple taxonomic levels and test different techniques for estimating times of divergence
4) Test different sampling strategies for reconstructing relationships within megadiverse taxa and assess the efficiency of competing analysis techniques in large-scale phylogenetics
5) Expand the 'Diptera Web' as THE central website for Diptera providing links to a comprehensive compilation of information on Diptera.

IV. Background: Overview of Dipteran Phylogenetics

The Diptera is one of the best-supported monophyletic orders of Insecta (Hennig, 1973: 37 autapomorphies). However, the identity of the sister-group remains controversial with Mecoptera, Siphonaptera (Hennig, 1973) or Strepsiptera being the best candidates (Whiting et al., 1997). For the past 40 years phylogenetic relationships within Diptera have been subject to intensive study (see Yeates & Wiegmann 1999) and this work has generated support for many traditionally recognized higher-level taxa and established major new clades (e.g., Eremoneura). However, the relationship between these remain unclear and dipterists are unable to replace the existing classification, which is riddled with paraphyletic groups, with a phylogenetic system. Traditionally three major groups are recognized of which two are paraphyletic:

(1) "Nematocera" The 'Nematocera' is a paraphyletic grade of approximately 35 families which contains more than one third of the known Diptera species. Several studies have examined the relationships among its main lineages. For example, Hennig (1973, 1981) recognized four infraorders: Tipulomorpha, Psychodomorpha, Culicomorpha, and Bibionomorpha (including Brachycera). Wood & Borkent (1989) provided the first post-Hennig cladistic hypothesis and partially resolved the relationships among the seven recognized infraorders (Tipulomorpha, Blephariceromorpha, Ptychopteromorpha, Culicomorpha, Psychodomorpha, Axymyiomorpha, Bibionomorpha). Several other phylogenetic studies followed (e.g., Courtney 1991; Sinclair 1992; Krzeminski 1992; Michelsen 1996), most focusing on particular subgroups or character systems, but the first and only published matrix appeared in Oosterbroek & Courtney (1995). This provided the foundation for subsequent studies of nematocerous flies (e.g., Friedrich & Tautz 1997) and its major subgroups (e.g., Pawlowski et al 1996, Miller et al. 1997, Saether 2000). Although the foregoing investigations provided many valuable insights on relationships, they resulted in distressingly few widely accepted hypotheses. Several major infraorders are well established, but the composition of 'Psychodomorpha', 'Tipulomorpha', and 'Bibionomorpha' is contentious, and the placement of families critical for reconstructing the origin of the brachycerous flies and, indeed, of all Diptera remains uncertain (e.g., Nymphomyiidae, Tipulidae, and Anisopodidae).

Key questions. Within "Nematocera" the key goals are (1) to resolve the relationships among the infraorders; (2) to test and rectify problems with the monophyly of several infraorders, especially the 'Bibionomorpha', 'Psychodomorpha', and 'Tipulomorpha'; and (3) to identify the sister-group of the Brachycera.

(2) "Orthorrhapha" The 19 families of "Orthorrhapha" comprise over 24,000 described species and occupy a critical, pivotal position between the "Nematocera" and the Cyclorrhapha. The relationships of the "Orthorrhapha" have been intensively studied over the past 30 years (Hennig, 1973; Woodley, 1989; Yeates, 2002). The characters are either discussed in the text, apomorphies are clearly associated with nodes, or an explicit data matrix is presented (Yeates, 2002). However, work to date is based on family-level groundplans, and character matrices for exemplar species are missing (but see Yeates, 1994). There is now general agreement on four infraorder-level clades (Stratiomyomorpha, Tabanomorpha, Xylophagomorpha, Muscomorpha, including Cyclorrhapha), but the relationships between the infraorders remain poorly supported by morphological, molecular, or combined data. This appears to be due to a lack of evidence rather than an unexpectedly high level of homoplasy which might point to a quick succession of brachyceran cladogenesis in the Jurassic (Wiegmann et al., in press).

The relationships of the Stratiomyomorpha are reasonably well supported (Sinclair et al., 1993) and there is also a modern hypothesis for the Tabanomorpha, although new questions were recently raised when the Pelecorhynchidae was treated as a rhagionid subfamily (Stuckenberg, 2001). The relationships within Xylophagomorpha have not been studied in a modern phylogenetic context, while those between the basal muscomorphan groups such as Nemestrinoidea and Asiloidea have come under detailed scrutiny. However, many questions remain. For example, the monophyly of the Nemestrinoidea, which occupy a critical, basal position in the Muscomorpha is still at best weakly supported. The relationships of the earliest speciose radiation of Muscomorpha, the Asiloidea, are poorly understood and there is growing evidence for a number of independent family-level lineages closely related to the Therevidae and Scenopinidae. Furthermore, while the muscomorphan clade Eremoneura is well established, relationships among basal groups of Empidoidea and Cyclorrhapha remain unclear (Collins & Wiegmann 2002a,b; Cumming & Sinclair 1995; Cumming & Sinclair 2000).

Key Questions Within "Orthorrhapha" the key goals are (1) to establish the relationships of the infraorders (2) to identify the sister-group of the Cyclorrhapha (3) to test the monophyly of the Nemestrinoidea and Asiloidea, (4), to reconstruct the phylogeny within Tabanomorpha, (5) to reconstruct the phylogeny within Asiloidea, and (6) to reconstruct the relationships among the important lineages of Empidoidea.

(3) Cyclorrhapha The approximately 90 recognized families of the Cyclorrhapha contain the largest proportion of the taxonomic and morphological diversity of Diptera. The prime sources for synthetic work are Hennig (1958; 1965; 1971; 1973; 1976), Griffiths (1972) and McAlpine (1989). The characters are discussed in the text, usually illustrated, and often mapped onto trees to indicate taxonomic distribution. However, matrices are missing and usually the work is based on groundplans. Three major groups are traditionally recognized. (1) The basal "Aschiza" are paraphyletic with the Syrphoidea probably being the sister-group of the Schizophora (=non-aschizan Cyclorrhapha). (2) Schizophora is well-supported and contains the bulk of the cyclorrhaphan families (>80). Traditionally two sections are recognized. The likely paraphyletic "Acalyptratae" contain approximately 65 families (e.g., Drosophilidae) and are the least understood section of Diptera. Resolving the relationships with morphology alone has proven difficult because the small body size of most species has apparently reduced morphological variability and caused much homoplasy. Furthermore, the large family-level diversity has impeded with synthetic work. The majority of families are small (40 less than 100 spp.; 54 less than 500 spp.), so relatively few exemplars are needed for adequate sampling. For many of the more speciose families sampling can be based on family-level cladograms (e.g., Drosophilidae, Tephritidae, Ephydridae, Agromyzidae). (3) The second and monophyletic section of the Schizophora are the Calyptratae which contains 14 families. Superfamilies have been recognized and are, with the exception of Muscoidea, probably monophyletic.

Key questions. Within Cyclorrhapha, the key goals are (1) to identify the most basal clades of Cyclorrhapha, Schizophora, and Calyptratae, (2) to find the sister-group of Calyptratae, (3) to test and rectify problems with the monophyly of some superfamilies (especially in 'Acalyptratae'), and (4) to resolve the relationships among the cyclorrhaphan superfamilies.

Available Molecular Evidence for Diptera. GenBank includes at least some sequences for standard "phylogenetic" genes (12S, 16S, 18S, 28S, COXI, COXII, EF-1a) for more than 300 species in 21 of the 35 "Nematocera" families, 25 of the 27 "Orthorrhapha" families and in excess of 500 species in more than 40 of the 90 Cyclorrhapha families. This creates an excellent basis for future work and in some groups all that is required is adding more sequence data and filling holes in the taxon sample. However, sampling across families is uneven with groups of medical, veterinary, and general biological significance having received most of the attention (e.g., Culicidae, Simuliidae, Phlebotominae, Therevidae, Tephritidae, Drosophilidae).

V. Proposed Phylogenetic Work

It is apparent from the overview that the main problem is the unknown relationships among major dipteran clades. The best approach to solving pressing phylogenetic problems in Diptera is thus reconstructing a solid backbone tree for these clades.

(1) Reconstructing this backbone tree is the goal of our first-tier analysis, which will be based on a small taxon sample of 42 species from all major dipteran lineages. The sample largely consists of model species in biology and fresh material is available from the tissue collections of the PIs. The inclusion of model species has the advantage that culture methods are known, much biological information is available, and their phylogenetic placement is of interest to a larger scientific community. The first-tier analysis will be based on the first comprehensive morphological data set for all of Diptera and a massive amount of DNA sequence obtained by sequencing 38 complete mitochondrial genomes and 30 kb of nuclear DNA sequence (15-20 genes).

(2) The larger taxon sample of the second-tier analyses is designed to test the monophyly of the major Diptera clades. Depending on support for monophyly and size, each of the 160 families will be represented by one or several species and the final matrix will comprise 240

species. A morphological data set will be assembled and combined with 12Kb (5-10 genes) of DNA sequence. The choice of genes will be guided by the results of the first tier analysis. Assembling the second tier data set will require much of the funding period and most superfamilies will initially be analyzed separately (see Management plan).

(3) The third-tier of the project relies on supertree techniques to build a cladogram for 1500-2000 species. The source cladograms will come from two initiatives. One is a comprehensive review of the phylogenetic Diptera literature of the past 40 years. Trees based on explicit character data (either in matrices or on cladograms) will be collected and recoded and/or integrated with the backbone tree reconstructed during our project. This highlights the importance of this tree because without this backbone no supertree could be generated for this megadiverse order. In addition we will offer four full PhD and six short-term fellowships for students who will provide cladograms for families and superfamilies requiring extensive taxon-sampling (e.g., Mycetophilidae s.l., Opomyzoidea, Calliphoridae). We estimate that our project will generate trees for more than 500 species of Diptera. We are aware of genetic information for 1200 species in Genbank and estimate that at least 1000 species are represented on morphology-based trees. Subtracting species overlap between analyses, we estimate that the Diptera supertree will connect 1500-2000 species by the end of the funding period.

VI. Molecular data: Choice of genes, methods and preliminary data

In our project we are targeting a large number and variety of genes from both the nuclear and mt genome. We will need a wide range of genes with varying substitutional dynamics, because we seek resolution at broadly differing levels of fly phylogeny, which exhibits a wide range of divergences (240-10 mya) and multiple rapid radiations. In choosing the genes, we can build on a growing knowledge-base on the genes utility (Caterino et al. 2000; Brower & DeSalle 1994). However, we will also need quantity, because there is growing evidence that analyses based on multiple genes often dramatically increase node support, even when high variability in a few genes provide substantial conflict (Olmstead et al. 1998; Baker & DeSalle 1999; Cognato & Vogler 2001; Kjer et al. 2001; Rokas et al 2001). While primary nucleotides, especially 3rd positions, for some of these genes may evolve too rapidly to be used individually, combined analyses employing appropriate models, or partition-based weighting schemes should allow us to both extract and experimentally characterize phylogenetic signal within our large sample.

Nuclear genes. These genes are useful, evolutionarily independent, phylogenetic markers, because of their location on different chromosomes. Using published or our own primer sequences, we will generate approx. 30 kb from the nuclear genome for the first tier species. To obtain this large a sample, we will begin with 15 kb based on established genes that are routinely obtained in our laboratories (e.g., 18S, 28S; EF-1a, PEPCK, DDC; e.g., Collins & Wiegmann 2002a; Friedrich & Tautz 1997). We will also draw on new genetic markers that were recently developed in the Wiegmann laboratory (CAD; 4 genes, 6.3 kb), and GART (3 genes, 4kb; see Results of NSF support). The remaining 15 kb will come from genes that are known for multiple dipterans (e.g, *Drosophila*, *Anopheles*, *Cochliomyia*), and have been successfully used in comparative studies of limited taxon range in Diptera or other insects (Bonacum et al. 2002; Caterino et al. 2000). These genes include *opsin rh1* (Mardulyn & Cameron, 1999), *wingless* (Baker et al. 2001), *white* (Krzywinski et al. 2001), and *period* (Regier et al. 1998). In developing these across the entire order we can draw on our strong collaborative ties to current insect phylogenomic projects (see Letters of support). Based on their performance, a subset of these genes will be selected for broader sampling at the second and third tier level.

Mitochondrial genes and genome. Animal mitochondria have a circular genome, which codes for two ribosomal RNAs (16S and 12S rRNA), 22 tRNAs and 13 proteins (cox1-3, nad1-6, nad4l, cob, atp6, atp8). Due to the ease of isolation, mitochondrial DNA has become a widely-used marker in molecular evolutionary studies (Caterino et al. 2000; Simon et al. 1994). While usually employed in phylogenetic studies at a relatively low taxonomic level, recent studies on vertebrate, deuterostome and arthropod relationships show that mtDNA is also phylogenetically informative at higher levels, when the conserved protein regions are concatenated (Castresana 2000; Hwang et al. 2001; Mindell et al. 1999; Waddell et al. 1999). The use of mitochondrial genome sequences is further enhanced by the occurrence of cladistically informative gene order rearrangement events (Boore & Brown 1998; Curole & Kocher 1999). Our preliminary analyses indicate that complete mitochondrial genome sequences might be as powerful a supplement to nuclear sequences for resolving dipteran phylogeny as it was for vertebrates and mammals (Springer et al. 2001). Diptera are an obvious test case for further study of the phylogenetic utility of mt genome sequences, because 8 of the 17 known insect genomes pertain to this order (Ballard 2000; Beard et al. 1993; Clary & Wolstenholme 1985; Lessinger et al. 2000; Lewis et al. 1995; Mitchell et al. 1993; Spanos et al. 2000). This also provides a good starting point for more comprehensive sampling in Diptera.

Preliminary data analyses: As a demonstration of the utility of combined data (rDNA & protein encoding genes) and mt genomes for resolving dipteran relationships, we analyzed the sequences for five genes from 38 of 42 first tier taxa spanning nearly 240 million years of fly evolution. The 3926 bp aligned preliminary data set includes 18S (508 bp), 28S (967 bp), 16S (399 bp) CAD (812 bp), and COI (1240 bp). Parsimony analysis in PAUP* 4.0 (heuristic search, TBR, 20 random taxon addition searches) with the 3rd positions of the protein encoding genes excluded (1375 variable; 842 informative sites) yields two equally parsimonius trees that differ only in the position of *Empis* (length=4663, CI=0.44 RI=0.38; Fig. 1; page 15). Despite encouraging agreement with expected relationships, it is apparent that some of the sequences included here may evolve too fast to be informative at deeper nodes. Better taxon sampling, additional genes, and morphological characters will be necessary for rectifying the problems.

We also analyzed a data set consisting of conserved mt protein (3586 aa sites) and ribosomal gene regions (1634 bp) for six dipteran and four outgroup species. The result gives consistent and well-supported resolution (fig 2; page 15). The screwworm fly (*Cochliomyia hominivorax*) groups with the drosophilid which supports the suspected paraphyly of the "Acalyptratae" (Griffiths, 1972; Beverly & Wilson 1984; Vossbrink & Friedman 1989; Yeates & Wiegmann 1999), which are here also represented by the fruitfly (*Ceratitis capitata*).

Molecular techniques. DNA extraction, PCR amplification, and sequencing of nuclear and mt genes are routine in the Wiegmann lab (e.g., Collins & Wiegmann, 2002a; 2002b). For PCR, we employ Panvera ExTaq polymerase and buffer following manufacturer protocols. For RT-PCR, we use the Perkin Elmer GeneAmpRNA PCR kit. This facilitates *in vitro* amplification of DNA coding regions by targeting transcribed RNA and producing cDNA copies. PCR products are run on agarose gels to confirm the size of amplified fragments, then cleaned using Qiagen Qiaquick mini-columns, sequenced in ABI dye-terminator cycle-sequencing reactions using Big Dye terminators, purified using Centri-sep column purification, and run on ABI 3700 Capillary Automated DNA Sequencers; (Macromolecular Core Facility of Wayne State University NCSU Genomics Research Lab; see Letters of Support).

The mt genomes will be sequenced in the Friedrich and Beckenbach laboratories where DNA extraction, PCR amplification and preparation of templates for sequencing of mtDNA are routine (Friedrich & Musquim, 2003; Stewart & Beckenbach, 2003). Automated sequencing is done in house at Wayne State, or using external sequencing services. The Friedrich and Beckenbach labs have developed a complete set of versatile PCR primers that allow the amplification of the entire coding region of mitochondrial genomes in short (ca. 1 kb) overlapping fragments for a wide range of insects. These primer sets will allow routine, high throughput sequencing of the complete coding regions of most dipteran taxa. Regions that prove more difficult will be completed using taxon specific primers based on partial sequence obtained using the standard primer sets. The control regions (A+T rich regions) will be determined where possible using taxon specific primers and either standard or long PCR techniques. Complete determination of control regions is not necessary for the major goals of this proposal.

VII. Morphological data: Bridging the gap between the suborders

Dipterists have traditionally worked in only one of the suborders, superfamilies, and/or families. This tradition has generated a confused morphological terminology with the non-homologous structures carrying the same name. All dipterists recognize that synthetic work is urgently needed for establishing homologies and for deriving character definitions that are useful across the order (Yeates & Wiegmann 1999). Few attempts have been made (Hennig, 1973; Michelsen, 1996; Wood & Borkent 1989; Sinclair et al. 1994; Cumming et al. 1995), but we are lacking a bold large-scale attempt that embraces the entire order and all major character systems. Such an initiative is proposed here. Three PIs on this proposal have recently published revised, quantitative morphological data matrices for the families of "Nematocera" (Oosterbroek & Courtney 1995), "Orthorrhapha" (Yeates 2002), and Lower Eremoneura (Wiegmann et al. 1993) that synthesizes information from different morphological character systems. These separate data matrices will be the starting point for a combined super matrix of the entire order.

The project team includes three morphology-based PIs representing the three Diptera suborders. These PIs and 13 additional collaborators representing Nematocera, Orthorrhapha, and Cyclorrhapha will carry out a morphological study of the 42 first-tier species. Upon completing, they will meet at a one-week workshop where homology problems will be debated, character definitions across the order will be established, and the morphological data matrix for the first-tier taxa will be built. Remaining homology problems will be discussed via email with additional collaborators and if morphology suggests alternative homology hypotheses, several codings will be prepared and tested phylogenetically (see Wiegmann et al., 1993; Meier & Hilger, 2000). The character definitions from the workshop will also form the basis for all further morphological work within the superfamilies and contribute the data for the anatomy atlas for the 'Diptera Biodiversity Web.' As detailed in the management plan, several superfamilies will be treated each year; and, for the larger taxa, separate cladistic analyses will be carried out.

The three PIs with strong morphology background will coordinate the work within their respective suborders involving collaborators, PhD students, and postdoctoral fellows (see Management plan). The PIs will ensure that the character definitions are consistent with the homology decisions taken at the workshop. In year 5 of the project a second workshop for all "morphological" contributors will be held. It is here that the remaining homology problems will be discussed and the final morphological data set for the second tier species is assembled. The morphology team currently consists of more than 10 professional dipterists, five postdoctoral fellows, and several PhD students thus ensuring that the each collaborator has ample time for a thorough morphological study of the 240 second-tier species and sufficient time for more specialized work on third-tier projects.

The morphological research will be based on light- and scanning electron microscopy. Wherever possible, characters from adults and from immature stages will be included (first-tier species). Morphological structures will be extensively documented using illustration techniques ranging from drawings over digital photos to SEM micrographs.

VIII. Challenges of large-scale phylogenetics and our contributions

Megadiverse arthropod clades pose one of the main challenges in assembling the tree of life. For example, it remains unclear which sampling strategy should be used. Should we collect a large amount of data for a small taxon sample (Murphy et al., 2001; Rosenberg and Kumar 2001) or use relatively few characters and a dense taxon sample (Zwickl & Hillis 2002; Hillis et al. 2003)? We are testing the first strategy in our first-tier and the second strategy in our second-tier analyses. The success of these approaches in Diptera will be judged by the strength of tree support, counting the number of nodes recovered in the analyses that are firmly established based on morphological evidence, compatibility with the fossil record, and the sensitivity of trees to different analysis conditions and weighting regimes. Should both strategies perform equally well, we will make recommendations for future studies based on a cost/benefit consideration.

Regardless of which sampling strategy is used for obtaining stable backbone cladograms, the construction of the tree of life for megadiverse taxa will probably have to partly rely on supertrees (Sanderson et al., 1998; Kennedy & Page, 2002), because groups like Diptera contain too many species to be analyzed simultaneously. However, the unconditional use of supertrees is currently inadvisable, because construction methods are still in their infancy and are untested for very large taxon samples (Bininda-Emonds & Sanderson, 2001; Salamin et al., 2002). Some construction techniques rely on consensus techniques and/or simply connect or combine source trees based on shared taxa (Sanderson et al., 1998; Wilkinson and Thorley 1998). These techniques are computationally inexpensive, but they struggle with conflicting source trees. The alternatives are techniques based on matrix representation parsimony (MRP) where trees are converted into matrices, which are then analyzed after combining.

In our project, we will explore the value of supertrees by addressing the following major issues (cf. Bininda-Emonds & Sanderson, 2001; Salamin et al., 2002). (1) We will compare the supertrees based on the source trees for the individual Diptera superfamilies/families reconstructed during the second- and third-tier analyses with the most parsimonious tree for the corresponding 'supermatrix' (Salamin et al., 2002; Sanderson et al., 1998). We will be able to establish whether supertrees are good estimators of trees based on a supermatrix and which techniques are most successful. (2) The matrices used in MRP contain a large proportion of missing data and it remains unclear for empirical data sets to what degree these missing values cause spurious resolution on supertrees (see Bininda-Emonds & Sanderson, 2001; Coddington and Scharff 1997). We will test whether supertree reconstruction using MRP is computationally feasible for matrices with a large number of taxa (1500-2000). This issue is largely unexplored because the data sets of previous supertree studies were small (<500 taxa). However, supertree techniques are only attractive if they can succeed to produce trees for large data sets. (3) An additional issue that we have already explored in a recent analysis is the use of implied weighting for analyzing MRP matrices (Goloboff, 1993a;b; see preliminary analysis below).

The large second tier character matrix with 240 taxa, several hundred morphological characters and 12Kb of DNA sequence data will also be an excellent data set for testing the computational efficiency of different analysis techniques, alignment strategies, and the power of different software packages. They can be used to test the efficiency of analysis strategies for large data sets (e.g. Parsimony Ratchet: Nixon 1999; Quicke, et al. 2001; EM algorithm for likelihood searching, J. Thorne, pers.comm.; Bayesian phylogeny estimation, Huelsenbeck and Ronquist 2001) and compare the performance of two different approaches to parallel computing ("dedicated cluster" versus “screen-saver” approach; see letter of support by M. Whiting). However, the main focus of our project is taxon-based and these analyses are intended to facilitate our main goal of reconstructing a solid tree for the Diptera.

IX. Diptera fossils and testing molecular dating techniques

Arguably, the fossil record of Diptera is more complete and better studied than the fossil record for any other insect order. Four periods can be distinguished. Most major nematocerous clades show significant diversity by the mid-Triassic (230 mya) and a Permian origin for the order can thus be postulated. The basal radiations of Brachycera appear in the Late Jurassic (160-145 mya) and the third major episode is the Cretaceous radiation of the Eremoneura (Grimaldi & Cumming, 1999) which are particularly well preserved in amber. The last period is characterized by the explosive Cenozoic radiation of schizophoran Cyclorrhapha. Most fossils are here "basal" members of extant families (e.g. Hennig, 1965) suggesting family origins in the Late Paleocene to Early Eocene (60-50 mya).

The excellent fossil record and morphological information especially on amber species provides the unique opportunity (1) to compare the chronological order of appearance of apomorphies with predictions based on phylogenetic analysis of extant species (Norrell & Novacek, 1992), (2) to study the impact of fossils on the results of cladistic analyses by including extinct species, and (3) to estimate the ages of major clades. As shown recently (Wiegmann et al., in press) the latter is particularly timely due to new Bayesian techniques for estimating divergence times (Thorne et al. 1998; Thorne & Kishino 2002). We will also be able to test the reliability of competing methods for dating by comparing estimates with ages based on the fossil record.

X. BioInformatics of Diptera

We propose to dramatically expand the existing www.diptera.org website into the 'Diptera Biodiversity Web' (DBW), THE central hub for information on Diptera in the worldwide web. The DBW will tightly link with the FlyBase for Drosophilidae and extend some of FlyBase's initiatives to all of Diptera (Gelbart et. al. 2002, see Letters of Support). The DBW will be home to (1) the world catalogue of Diptera 'BioSystematic Database of World Diptera (BDWD)' with a new species interface and internet portal; (2) a newly created 'World Stock List of Diptera' with overview over which dipteran species are cultured in biological and medical laboratories around the world; (3) an expanded Dipterists' Directory (http://www.diptera.org/worlddip.htm) including data on past workers with and information on the whereabouts of their collections, papers and field notes; (4) interactive keys; (5) an anatomy atlas or MorphoBase similar to the atlas for Drosophilidae in FlyBase. This atlas is based on the morphological research of the project and provides multiple synonymous names for same structures. It helps to overcome problems with interpreting old literature; (6) the most current supertree for Diptera; (7) a searchable database that links DNA sequences with identification information on the specimen from which the DNA was obtained (where possible). Due to frequent identification problems such a link is desirable for invertebrates; (8) a query tool similar to the Species Analyst to consolidate information on species holdings of various museums. The DBW site will also link to the updated Tree-of-Life pages for Diptera with its background information on the various taxa (see Letters of Support, David Maddison), and to all major collection inventories (general collections and type holdings).

XI.1. Project participants and project coordination

Phylogenetic component. The proposd project requires considerable morphological and taxonomic expertise in order to cover the full diversity of Diptera, selecting an appropriate taxon sample, and to coordinate collecting. Therefore, each suborder is represented by a morphologist (G. Courtney: 'Nematocera;' D. Yeates: 'Orthorrhapha;' R. Meier: Cyclorrhapha) who has recruited a team of additional collaborators: A. Borkent, D. Amorim, J. Skevington, D. Grimaldi, S. Gaimari, L. Papp, S. Marshall, T. Pape, H. de Jong, P. Adler, M. Irwin, C. Lambkin, S. Winterton. The molecular work requires facilities with a proven ability in generating DNA sequences. The Beckenbach, Friedrich and Wiegmann labs fulfill this requirement with the Beckenbach and Friedrich labs being experienced with generating and analyzing large-scale mitochondrial sequences (Friedrich & Mosquim2003; Stewart & Beckenbach 2003), and the Wiegmann lab having a decade of experience with sequencing many different genes for a broad array of Diptera. Molecular datasets will also be generated in three additional collaborating laboratories: S. Scheffer (USDA-ARS, Systematic Entomology Lab, see letter of support); J. Skevington (CA Dept. Ag.; see Letter of support); co-PI R. Meier (National University of Singapore). All PIs have a proven track-record in systematics and are intimately familiar with up-to-date analysis techniques in phylogenetics. They also routinely collaborate on projects and have jointly organized conference symposia.

Bioinformatics. The proposed work also requires an experienced team of bioinformatics experts. Over the past eight years, Gail Kampmeier designed, expanded, and maintained the NSF PEET-sponsored database system 'Mandala'. This system, designed for cross-platform use in FileMaker* Pro, currently manages specimen, nomenclatural, and literature information for systematics and biodiversity projects (http://pherocera.inhs.uiuc.edu/MandalaModel.pdf) (http://pherocera.inhs.uiuc.edu);(http://www.inhs.uiuc.edu/inhsreports/autumn-01/fblitz.html).

Mandala is used to construct reports, publications, distribution maps, and linkages to the internet (GenBank, collections webpages, electronic publications, etc.). We will use Mandala to track the specimens from which DNA is extracted, provide links to GenBank URLs, and will expand the functionality of Mandala to include character data, phylogenetic data and images. F. Christian Thompson has decades of experience with building databases and is the collaborator in charge of the existing Diptera website (www.diptera.org) as well as a leader in many International efforts (International Commission on Zoological Nomenclature, Species2000, ITIS, CICD). The BDW will be hosted by the server of the US Department of Agriculture and an international steering committee will be established in the first year of the project. Chris Thompson will co-supervise a postdoctoral fellow who will populate information in the greatly expanded databases and web pages. The fellow will also contribute to the species interface and internet portal for the BDWD. The phylogenetic literature for Diptera will be reviewed by the PIs and the trees and data matrices sent to NC State University for processing. Here the trees will be converted into the tree-of-life format and the character matrices digitized by undergraduate students. The data will then be submitted to the Tree-of-Life and TreeBase databases (B. Wiegmann and D. Yeates, two of the Co-PIs, are co-editors of the Diptera pages of TOL).

XI.2. Taxon sampling, vouchering, techniques

The taxon sample for the first tier reflects two demands: (1) We include one species for each major clade; (2) we also include as many model species as compatible with the phylogenetic needs in order to satisfy the interests of the larger biology community. The second tier analysis samples all families (at least for morphology). Families that are either large or whose monophyly is contentious are sampled with multiple species. The species are selected by taxonomic experts to ensure that all critical clades are included. Many experts are official collaborators or have agreed to informal collaboration (see Letters of support). Getting material for the exemplar species is challenging, but the PIs routinely collect and have assembled an excellent group of dipterists to cover the taxonomic and geographic range of Diptera. Africa: D. Barraclough (South Africa), M. Irwin (faunal projects in Madagascar, Fiji, Australia, Chile), Southeast Asia: R. Meier (Malay Peninsula), G. Courtney (Thailand), Palaearctic material: B. Merz, L. Papp; H. de Jong, J. Gelhaus (Mongolia); Australia: D. Yeates; Nearctic: A. Borkent, D. Grimaldi, B. Wiegmann, C. Thompson, G. Courtney, S. Marshall; Neotropical: A. Borkent, S. Marshall, D. Amorim. Voucher specimens will be deposited at the Insect collection of the NC State University. Duplicates will be sent to the Smithsonian and the American Museum of Natural History. The tissue collection of the AMNH will also become the depository of duplicate tissue samples.

XI.3. Phylogenetic methods and analytical hypothesis testing

FIRST TIER. The relative small number of taxa in the first tier data set makes it an ideal matrix for experimental work on different analysis techniques. (1) Alignment. The data set will contain a large number of ribosomal genes and we will continue with our comparative work on different alignment strategies (e.g., Meier & Wiegmann, 2002; Friedrich & Tautz 1997; manual vs multiple-sequence and optimization alignment; treatment of gaps as missing data or fifth character state; secondary structure based alignments). Given that we can generate a stable tree based on the large amount of alignment-independent data we will be able to assess which alignment strategy maximizes congruence. With the exception of the optimization alignment analyses carried out with POY, we will use PAUP*4.0 (Swofford 1999) and the Parsimony Ratchet of NONA implemented in WinClada (Nixon) for this work. (2) Sensitivity analysis. We will compare cladograms obtained under a large number of different weighting regimes for gap-costs, 3rd positions in protein-coding genes, and transformation models in order to determine those analysis conditions that maximize congruence with the morphological data (Wheeler, 1996, see also Meier & Wiegmann, 2002). (3) Analysis philosophies. We will compare the trees obtained under the three main analysis philosophies, parsimony, maximum likelihood, and

Bayesian likelihood while seeking to characterize fit between data and tree estimates. We will also explore the feasibility of scoring morphological data to obtain character sets that could be analysed using recently proposed likelihood methods (Lewis 2001). (4) Gene utility. The data set will provide excellent raw material for comparing the relative utility of morphology, genes, and genomes at different taxonomic levels, and assessing the relative support from the different data partitions (e.g., Meier & Wiegmann, 2002). Tree support will be assessed using standard techniques such as bootstrap and jackknife analyses as well as Bremer values.

SECOND TIER. The complete second tier data set is assembled superfamily by superfamily over the period of several years. The superfamily-level character matrices can be analyzed using the same techniques as described for the first-tier analysis. However, the analysis of the full second-tier data set will require different computational techniques due to the large number of terminals (240). Here we will pursue the same analysis goals as for tier 1, but compare the relative efficiency of analyzing large data sets using specialized software like TNT with the use of parallel computing. A rackable system of parallelized computers and supercomputing facilities for phylogenetic computation are available to the project from the NC State University Bioinformatics Research Center and the North Carolina Supercomputing Facility (see Management plan, Letters of support). For analyses requiring PAUP* we will also use the screensaver approach developed by M. Whiting at Brigham Young University.

THIRD TIER: SUPERTREE CONSTRUCTION. We have recently reconstructed the first supertree for all 151 Dipteran families. The superfamily-level semi-strict consensus tree based on 1879 trees is shown in Figure 3. It is based on a 313-character MRP matrix (Baum 1992, Ragan 1992) derived from 12 primary source trees (Griffiths 1972, Hennig 1973, Wood and Borkent 1989, Woodley 1989, McAlpine 1989, Pape 1992 Sinclair et al. 1994, Cumming et al. 1995, Oosterbroek & Courtney 1995, Matile 1997, Yeates 2002, Collins & Wiegmann 2002). A heuristic parsimony analysis was conducted with PAUP* 4.0b10 using Goloboff's weighting function and an all zero outgroup (10 random addition sequences, NNI branch swapping; k ranging from 1 to 8). For the third tier analysis we will continue to use different implementations of 'matrix representation parsimony' (MRP; cf. Salamin et al, 2002; Sanderson et al., 1998). We will convert the trees into matrices and then combine the data sets (e.g., using SuperTree: Salamin et al, 2002; TreeView: Page, 1996; and RADCON: Thorley & Page, 2000). These matrices are then analyzed using general parsimony but we will also continue to experiment with implied weighting (Goloboff, 1993a;b) and weighting schemes that assign higher weights to source-tree nodes with high branch support (see Sanderson et al, 1998).

XI.4. BioInformatics products and protocols

Today there is a wide range of initiatives seeking to expand our knowledge of BioDiversity (GBIF, ITIS, Species2000, Discover Life, NBII, Species Analyst, EcoPort, etc.). We seek to provide a taxonomic focus for 10% of this biodiversity with our Diptera Web site and we will strive to coordinate our work with all these effort and to make our information conform to community standards. The BioSystematic Database of World Diptera has, and continues to provide, name information to ITIS and Species2000 (about 10% of their names came from BDWD or its earlier incarnations). The 5 year course of this project should allow the majority of the information in the BDWD validated by peer-review. The species interface and internet portal for the BDWD will be a means of finding information by common user's requirements, such as conservation status (endangered, threatened, etc.), environmental and agricultural significance (endemic, alien, invasive, pest, pollinator, biological control agent). MANDALA will be enhanced so its specimen information can be integrated with that of other organisms via the Species Analyst project (tsadev.speciesanalyst.net). A tool similar to Species Analyst will be used to perform queries and consolidate the results from most of the available species inventories of the major museums of the World. The K-12 features, such as the Diptera Safari, will be expanded with more pictures and information and additional education features will be added.

XI.5. Outreach and education

Graduate- and undergraduate- level training directly funded by the project will be conducted at North Carolina State, Iowa State, Simon Fraser, and Wayne State Universities.

Postdoctoral Training - Five postdoctoral fellowships are requested. These positions will be critical to providing new systematics PhD graduates (many generated by PEET initiatives in Diptera) employment and training opportunities while building their professional careers. The commitment to postdoctoral training insures that the project is a catalyst for generating further research opportunities in the dipterological systematics community.

Graduate training. We will fund four PhD students through the project and provide summer salary for a 5th at Wayne State University - these students will complete dissertation projects in dipteran phylogenetics under the supervision of the specific PI. All of the projects will involve molecular and morphological data collection and phylogenetic analyses. In year 1, the students will be recruited and enrolled in PhD programs in Entomology (NCSU, IASU) and Biological Sciences (Simon Fraser, Wayne St.), student committees will be selected from the faculty, and projects will be designed and planned with the appropriate PI and circulated among project participants. In each of years 1-5, graduate students will complete course requirements, with concentrations in systematics, molecular evolution, genomics, and evolutionary biology. Students will be trained in systematics methods and be given opportunities to visit project participants or taxonomic specialists for extended training, and to attend project group meetings. Each student will present papers or posters at 2-4 annual meetings of ESA or SSB.

Graduate internships: Six summer research internships for graduate students will be offered to US or international students working on graduate degrees in Diptera phylogenetics. Students will be trained in molecular systematics in the Wiegmann Lab at NCSU and learn how to acquire and process molecular data, and how to analyze the data. Interns will contribute source trees for the 3rd tier analysis.

Undergraduate training. In each of the 5 years, an undergraduate Biological Sciences major at NCSU will conduct a research internship in the Wiegmann laboratory. These students will learn molecular phylogenetic techniques, collect data for 1 or 2 genes for a small sample of taxa, participate in lab meetings, and learn aspects of fly biology and classification. These students will present their research findings at the Annual NCSU Undergraduate Research Symposium. In each of years 2, 3 and 4, a summer undergraduate intern in the Wiegmann lab will come from the cooperative training programs between NCSU and NC A&T University, a program that provides research mentoring and internships for minority science majors.

In each year of the grant students enrolled at the National University of Singapore and the senior undergraduate subject at the Australian National University (BIOL3115) will contribute to the project by carrying out small research projects.

Outreach. The Diptera Web (www.diptera.org) will be used throughout the project to disseminate information about Diptera to both research and educational institutions and the general public. The website already includes educational components for K-12 students and teachers (e.g, Diptera Safari, Dr. Seuss on Ann Anopheles) and has prominent links to other websites about Diptera. We will actively develop the K-12 educational offerings and provide links and revisions to the Diptera sections of the "Tree of Life" website (tolweb.org; see Letters of support). Project status reports and published products will be added to the website. PI Yeates will coordinate the development of a series of web pages and associated posters that provide an illustrated identification guide to the major groups of Diptera and their diversity and biology. Yeates is currently developing an interactive key to adult Diptera families in the software platform Lucid. This project will be developed in parallel with the Tree of Life for Diptera. The PIs will use these products to develop portable exhibits on Diptera biology, diversity, and phylogeny for public display at BUGFEST events in local natural history museums (e.g., North Carolina Museum of Natural Sciences; Smithsonian Institution; Raffles Biodiversity Museum, Singapore; Illinois Natural History Survey Mobile Science Center).

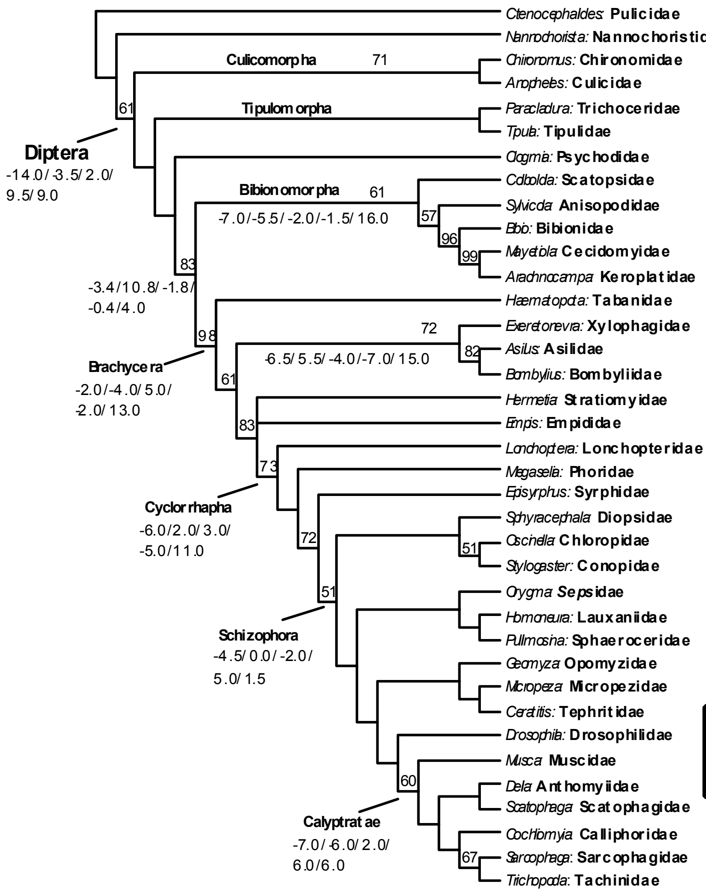

Ctenocephaldes: **Pulicidae**
Nannochorista: **Nannochoristidae**
Culicomorpha
71
Chironomus: **Chironomidae**
Anopheles: **Culicidae**
61
Tipulomorpha
Paracladura: **Trichoceridae**
Tipula: **Tipulidae**
Diptera
-14.0/-3.5/2.0/
9.5/9.0
Clogmia: **Psychodidae**
Bibionomorpha
61
Colbolda: **Scatopsidae**
Sylvicola **Anisopodidae**
-7.0/-5.5/-2.0/-1.5/16.0
57
Bibio: **Bibionidae**
96
Mayetiola **Cecidomyidae**
99
Arachnocampa **Keroplatidae**
83
-3.4/10.8/-1.8/
-0.4/4.0
Haematopota: **Tabanidae**
72
Exeretonevra: **Xylophagidae**
98
Asilus: **Asilidae**
82
-6.5/5.5/-4.0/-7.0/15.0
Brachycera
Bombylius: **Bombyliidae**
61
-2.0/-4.0/5.0/
-2.0/13.0
Hermetia **Stratiomyidae**
83
Empis: **Empididae**
Lonchoptera: **Lonchopteridae**
73
Megaselia: **Phoridae**
Episyrphus: **Syrphidae**
Cyclorrhapha
-6.0/2.0/3.0/
-5.0/11.0
Sphyracephala **Diopsidae**
72
Oscinella **Chloropidae**
51
Stylogaster: **Conopidae**
51
Orygma **Sepsidae**
Homoneura: **Lauxaniidae**
Schizophora
-4.5/0.0/-2.0/
5.0/1.5
Pullimosina: **Sphaeroceridae**
Geomyza **Opomyzidae**
Micropeza **Micropezidae**
Ceratitis: **Tephritidae**
Drosophila: **Drosophilidae**
Musca **Muscidae**
60
Delia **Anthomyiidae**
Scatophaga **Scatophagidae**
Calyptratae
-7.0/-6.0/2.0/
6.0/6.0
Cochliomyia **Calliphoridae**
67
Sarcophaga: **Sarcophagidae**
Trichopoda: **Tachinidae**

Figure 1(left) Strict consensus of two mpts for the 3.9 Kb combined molecular data (3rd position sites removed) for 37 tier 1 taxa. Values above the node are bootstrap percentages above 50% for 500 replicate searches. Values below the node are paritioned Bremer support values (16S/ COI/18S/28S/CAD)

Figure 3 (below). Dipteran Supertree produced by MRPcoding of all major phylogenetic hypotheses produced in the Diptera over the past 30 years. The data was coded at the family level, and this supertree is a summary of the full supertree at Infraorder / Superfamily level. See text for details, and http :// www.inhs.uiuc.edu / cee / therevid / classif.html.

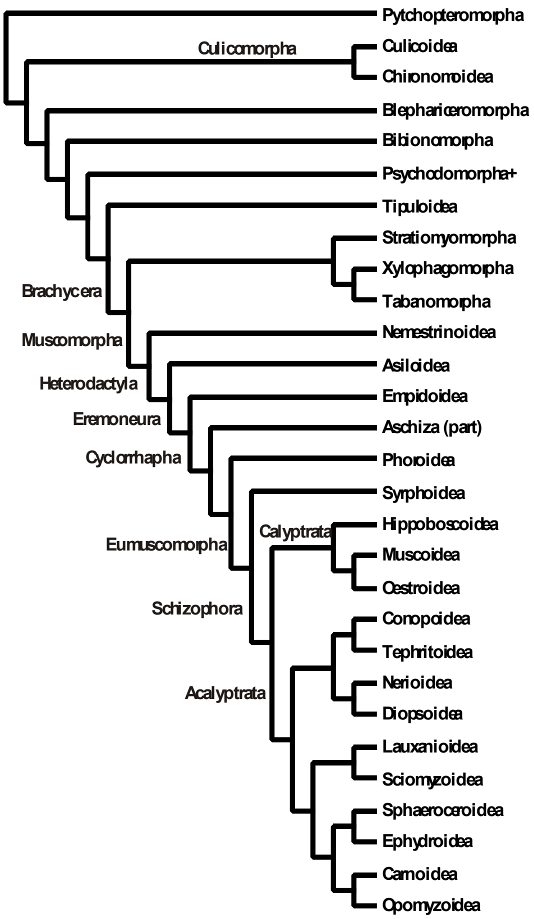
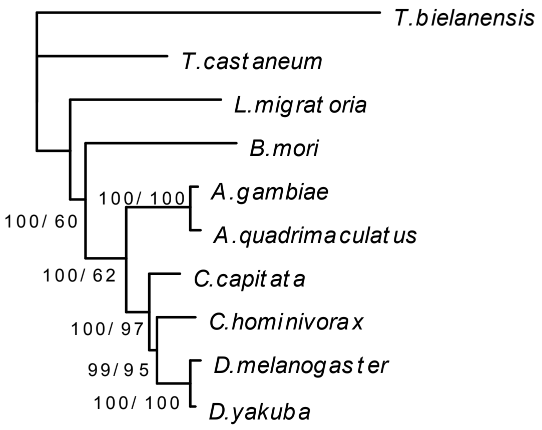


Figure 2: (below) Maximum likelihood tree based on dipteran combined conserved mitochondrial protein regions (left branch support bootstrap value) or mitochondrial rRNA sequences (right branch support bootstrap value).


T.bielanensis
T.castaneum
L.migratoria
B.mori
100/100
A.gambiae
100/60
A.quadrimaculatus
100/62
C.capitata
100/97
C.hominivorax
99/95
D.melanogaster
100/100
D.yakuba