1-1-2012

Phylogenetic utility of mitochondrial and nuclear genes: a case study in the diptera (true flies)

Jason Caravas
Wayne State University,

PHYLOGENETIC UTILITY OF MITOCHONDRIAL AND NUCLEAR GENES: A CASE STUDY IN THE DIPTERA (TRUE FLIES)

by

JASON CARAVAS

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2012

MAJOR: BIOLOGICAL SCIENCES

Approved by:

_____

Advisor Date

_____

_____

_____

DEDICATION

I would like to dedicate this work to my family. They have all taught me so much that it is impossible to do them justice here, but here is the place to try. From my father Gary, I learned how to make people laugh and the nearly lost art of proper spelling ("TAR"-"ANT"-"ULA"). From my mother Victoria I learned how to manage a menagerie of exotic critters. I doubt she has any idea how often I have discussed the breeding habits of Madagascar hissing beetles or the proper care and feeding of giant millipedes with my colleagues, but these topics have come up more than once. From my brother Justin, I learned to share, although as twins I'm not sure we ever had much choice. He has been my peacemaker and mediator for my entire life. From my brother Bradley, I learned to follow my dreams and that Frisbee golf was an actual sport that some people take seriously. And from my grandmother Marie "Sidecar Granny" Kurz, I learned that in the end, the only thing that really matters is enjoying oneself. Without the support of my family, I would not be here today. Thank you all. I could ask for no better family.

Of course, I'd be remiss if I did not also mention Dr. Bethany Strunk, who has been my confidant and companion for nearly my entire tenure in graduate school. She somehow managed to graduate before me, but I have (almost) forgiven her for showing me up. Thank you for putting up with me.

I'd also like to include a dedication to Frank and Cathy Strunk. I appreciate how you have always made me feel welcome in your home despite my somewhat unusual manner. Frank, I enjoy our discussions on politics and your pop quizzes on literature and history, irregardless of your occasional grammatical quirk. Thank you.

ACKNOWLEDGEMENTS

I would like to thank my committee members for their guidance and forbearance of my perpetually shifting graduation date. I would like to extend special thanks to Dr. Vipin Chaudhary, whose encouragement in to learn programming and grid computing had a dramatic impact on the direction of my research

I would like to thank my mentor Dr. Markus Friedrich for all of his help. I never doubted that his primary goal was to prepare me for a successful career in science rather than merely using me to further his own personal research. He went out of his way to encourage me to attend every workshop or conference that was even remotely relevant to my interests and to apply for every funding opportunity or summer project that would add to my C.V. In the lab, I had free reign to explore my interests and he frequently offered me side projects that he knew I would enjoy. The breadth of my experiences in his lab far exceeds the scope of this dissertation and I am grateful for the opportunity to graduate with such a well developed skill set.

My time in Dr. Friedrich's lab would not have been nearly as enjoyable without the great people I have worked with. Dr. Ying Dong, Dr. Zhenyi Liu, Dr. Xiaoyun Yang, Suchitra Balasubramanian, Mitchell Walker, Magdalena Jackowska, Nazanin Zarinkamar, Meng Wu, Riyue Bao, Anura Shrivastava, Zahabiya Husain, Dr. Arun Sasikala-Appukuttan, and Qing Luan have all been great friends to me.

Lastly, but not least, I would like to acknowledge the help of my undergraduate and high school research assistants who helped with mitochondrial genome sequencing. Ivanna Yavorenko, Robert Hanrahan, Mithun Neral, Jessiedeep Multani, and Farvah Fatima all contributed to this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 “INTRODUCTION”

Mitochondrial and nuclear gene biology

Within animal cells, two independent genomes with different modes of generational transmission coexist. The nuclear genome, found within the nucleus of the cell, is inherited from both parents in sexual organisms. In most diploid cells, two copies of each gene (one from each parent) are present. The vast majority of genes are found in the nuclear genome of the cell, with most eukaryotic organisms having between approximately 5,000 (Wood et al. 2002) and 28,000 (Jaillon et al. 2004) genes.

The other genome is found in the mitochondria, a cellular organelle. These small, circular genomes are maternally inherited and present in multiple copies in each mitochondrion. Most cells contain dozens or even hundreds of copies of mitochondria, thus the mitochondrial genome has a much higher copy number than the nuclear genome. In contrast to the large nuclear genome, the mitochondrial genome of animals typically contains a highly conserved set of only 37 genes with the majority (22) being short transfer RNAs (Boore 1999), although mitochondrial genome content can differ dramatically outside of the animals (Burger, Gray, Franz Lang 2003).

As the two genomes are found in radically different cellular compartments, the mutational forces acting upon them are equally distinct. It is difficult to generalize the evolutionary constraints acting on each of the very diverse set of nuclear encoded genes however their environment and mode of replication can be characterized. The nuclear genome is packaged and protected by histones, reducing the availability of bases to participate in chemical interactions that might result in a substitution (Enright, Miller, Hebbel 1992; Ljungman, Hanawalt 1992). Proofreading activity in the nucleus during

and between replication corrects substitutions, and a second copy of each chromosome exists in diploid organisms which can be used as a template for repair through homologous recombination. The high degree of protection and proofreading fidelity of the nuclear genome tends to result in a very slow rate of substitution in nuclear encoded genes with purifying selection acting to further limit the rate of change. The mitochondrial genome, on the other hand, is not packaged as chromatin and is thus more exposed to the mutagenic free radicals that are produced in the mitochondria as a result of respiration. While base excision repair mechanisms are known to be functional within the mitochondrion and recent evidence suggests that other nuclear repair mechanisms may also be functional, these repair mechanisms likely only serve to mitigate the rate of mitochondrial DNA damage rather than prevent or reverse it (Gredilla, Bohr, Stevnsner 2010). Furthermore, mitochondrial replication is believed to take a relatively long period of time compared to nuclear genome replication, leaving one strand as more vulnerable single stranded DNA for an extended period (Clayton 1982; Bowmaker et al. 2003). This results in a long term bias towards adenine and thymine in mitochondrial sequences due to deamination of cytosine to uracil in the lagging strand. Lastly, proofreading during replication of the mitochondrial genome has been found to be inefficient in some mammalian cells due to biases in the mitochondrial dNTP pool (Song et al. 2005). In sum, these differences typically result in an increased rate of substitution in mitochondrial genes estimated to be 4.5 to 9 times faster than the rate of substitution in an average nuclear gene in Drosophila (Moriyama, Powell 1997) and even higher in other groups (Brown, George, Wilson 1979; Oliveira et al. 2008).

Mitochondrial and nuclear genes in phylogeny reconstruction

At first glance, nuclear encoded genes would seem to be to an obviously superior source of phylogenetic information, especially for more ancient divergences where multiple substitutions at variable sites can lead to the obliteration of phylogenetic signal. From a practical point of view, however, mitochondrial genomes have much to recommend them. The high copy number of the mitochondrial genome relative to the nuclear genome makes amplification of mitochondrial gene fragments by polymerase chain reaction (PCR) an easier task than the amplification of nuclear genes. Furthermore, while variations in mitochondrial sequence can and do exist in the same organism and even the same cell due to their high mutation rate, they share a great deal of sequence similarity due to fact that they are all descended from a single small population of mitochondria inherited from the maternal parent. In contrast, nuclear genes can exist in two distinct variations (alleles) on the paternal and maternal chromosomes, complicating the amplification of a single sequence and determining which allele to use in phylogenetic reconstruction. As the gene order on mitochondrial genomes is typically conserved and rearrangements must take place in the context of a small (<20 kb) circular genome with a very small amount of noncoding sequence, amplification of full gene sequences or multiple gene sequences is trivial. This allows for efficient recovery of sequence data from poorly preserved biological samples, such as feces or ancient DNA, where the long, low copy strands of nuclear DNA may be too fragmented to amplify. Nuclear genes are typically found spread out throughout the chromosomes with large non-conserved intergenic regions between them. Thus, nuclear genes must be amplified

from conserved internal motifs and amplification of the entire gene can be challenging. Mitochondrial genes also lack non-coding introns and can be sequenced from one end to the other. Nuclear genes which contain introns must be sequenced in pieces. Lastly, despite a high rate of substitution at variable sites, the mitochondrial genes all play crucial roles in cellular respiration and have many regions under strong purifying selection, resulting in blocks of highly conserved sequence which can be targeted with PCR primers (Simon et al. 1994; Castresana 2000). Depending on the exact function and evolutionary constraints acting on a particular nuclear gene, regions of high variability may be present which complicate amplification and alignment of the gene.

Both mitochondrial and nuclear encoded genes have been used with great success for phylogenetic inference. Small sets of nuclear genes first produced trees which unified the crustacean and hexapods into Pancrustacea (Friedrich, Tautz 1995) and cast doubt upon Articulata (the traditional placement of annelids as the sister group to arthropods) by proposing the radically unorthodox Ecdysozoa clade (Aguinaldo et al. 1997). Mitochondrial gene phylogenies have provided early insights into mammalian and avian evolution (Mindell et al. 1999; Waddell et al. 1999), deuterostome divergences (Castresana et al. 1998), have proven informative on ancient arthropod divergences (Hwang et al. 2001), and have suggested reconsideration of chordate relationships (Zhong et al. 2009). Disagreement between phylogenies derived from mitochondrial and nuclear genome sources are not uncommon (Galewski et al. 2006; Zink, Barrowclough 2008), however these disagreements can often be resolved with alternative methods or appropriate treatment of mitochondrial gene data (Gibson et al. 2005; Hassanin 2006).

Despite the proven performance of mitochondrial gene phylogenies, the more common failures of mitochondrial gene trees to resolve phylogenetic questions has caused their utility to come into question (Lin, Danforth 2004; Zink, Barrowclough 2008). As the underlying mechanisms of mitochondrial evolution suggest that mitochondrial sequence should be less informative than nuclear gene data on a per site basis due to decreased sequence complexity (AT bias) and saturation (multiple substitutions), it is unsurprising that mitochondrial data performs poorly when compared to nuclear gene data in a per site manner. The fact that many mitochondrial gene trees rely on only a small subset of available mitochondrial genes compounds the problem by not compensating for decreased per site informativeness with an increased number of sites. Modern model based phylogenetic methods are statistically consistent (as the amount of sequence data increases towards infinity, the probability of producing the correct topology approaches 1.0) (Fisher 1922), therefore sampling a greater number of mitochondrial genes could dramatically increase the performance of mitochondrial gene phylogenies. The performance of larger amounts of mitochondrial gene data (up to the full mitochondrial gene complement) may provide a level of phylogenetic utility greater than is suggested by its per site performance. Rigorous testing of complete mitochondrial sequence data against comparably sized nuclear gene data sets is an area that requires further exploration.

Divergence time estimation

Phylogenetic tree inference methods rely on the assumption that substitutions accumulate over time in related sequences. Consequently, very similar sequences are likely to be closely related as few substitutions have occurred in each sequence. Model

based inference methods attempt to model sequence evolution in a more nuanced way than merely counting substitutions, however sequence similarity still plays a large role in determining phylogenetic relatedness.

In tree reconstruction, the time dimension of the evolutionary process is often discarded as a nuisance parameter and a more abstract measure of substitutions per site is used to measure how closely related the sequences are. With external information about the rate of substitution accumulation in the sequences of interest, the time dimension can be estimated and the date of sequence divergence and the age of their most recent common ancestor (MRCA) can be estimated. For species tree reconstruction, the external information on the rate of substitution is typically provided by dated fossils believed to represent minimum or maximum ages for clades in the tree.

The earliest attempts at molecular divergence time estimation assumed a global clock (a constant rate of substitution) applied to all sequences at all time points in the tree (Zuckerkandl, Pauling 1962; Margoliash 1963; Zuckerkandl, Pauling 1965; Sarich, Wilson 1967b). This simplifying assumption allowed any node on the tree to be dated with a single calibration point as all genes sequences were assumed to accumulate substitutions at the same rate. "Clock-like" genes which did not violate this assumption were uncommon (Goodman 1981a; Goodman 1981b; Czelusniak et al. 1982), possibly non-existent, thus global clock methods were replaced with local clock methods when they became available (Yoder, Yang 2000; Douzery et al. 2003; Aris-Brosou 2007; Svennblad 2008; Drummond, Suchard 2010). Local clock methods assume that the rate of substitution can vary across the tree but that related clades or sequences are likely to share a similar rate of substitution (a clock) and that related clocks are likely to be

similar. Because multiple rates are assumed, multiple fossil calibration points can be used throughout the tree to assist in assigning clocks to nodes.

Since the advent of local clock models, divergence time estimation has become a common corollary to phylogenetic studies. As the fossil history is incomplete, the true ages of MRCAs are usually totally unknown, and local clocks represent a simplification of a poorly characterized process, these divergence time estimates represent best guesses as to clade ages and are difficult to verify. Further confusing the issue, there has been little work regarding the appropriate data sources or preparation techniques for divergence time estimation. As a result, most divergence time estimates are the result of *ad hoc* methods which use whatever data is conveniently available. No data exists on whether mitochondrial or nuclear genes give different results or whether the inclusion of highly variable third codon positions or variable gene regions has an impact on inferred ages. As no standards of data preparation for divergence time estimation have been rigorously tested, this represents an open question in need of study.

Dipteran diversification: a Gordian superknot on wings

The insect order Diptera ("true flies") is well established as a monophyletic group with clearly recognizable synapomorphies (shared derived characters). Perhaps the most recognizable synapomorphy of the group is the reduction of the hind wings to club like balancing organs known as halteres. The halteres gyrate to stabilize the fly in flight, allowing precise control of pitch and roll as well as hovering. Due to the presence of halteres and the powerful flight muscles in the mesothorax, dipterans are some of the most nimble and adept fliers of the insects.

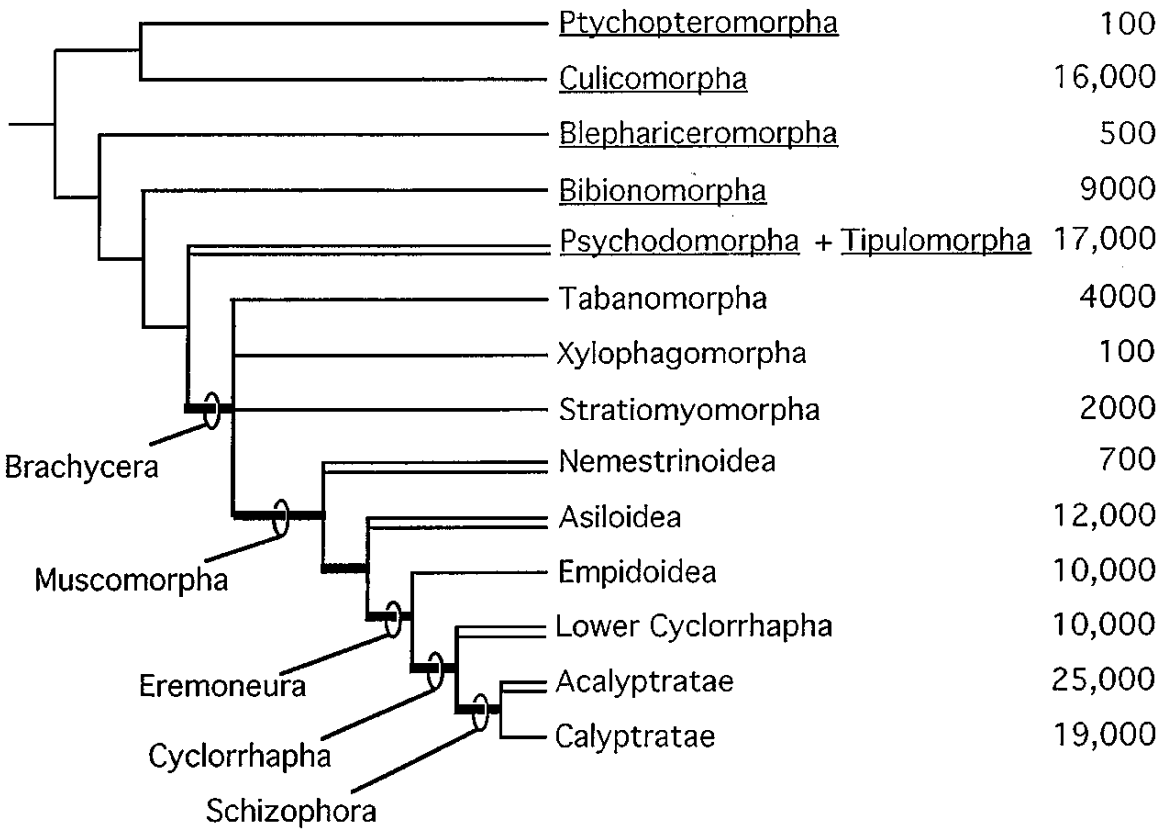
Similar in age to Coleoptera (beetles), the dipterans represent one of the four major holometabolous lineages, along with the Hymenoptera (ants, bees, and wasps) and Lepidoptera (butterflies and moths). The Diptera are the second most species diverse animal group after their cousins the Coleoptera. The megadiverse Diptera includes more than 150,000 described species (Pape, Thompson 2010) and accounts for approximately 12% of known animal species.

Fossils of the four winged family Permotipulidae, a stem-group of the Diptera with reduced hindwings and an enlarged mesothorax, date back to the Upper Permian (250 million years ago) (Willman 1989). The earliest true fly fossil dates to the mid-Triassic, placing a minimal age of 240 million years on the order Diptera (Krzemiński 2003). Primitive dipteran lineages are present in fossils from the Upper Triassic, with the a large proportion of fly fossils dating to the Mesozoic (Hennig 1981; Evenhuis 1994; Labandeira 1994).

The order Diptera is traditionally divided into two suborders: the Nematocera (long-horned flies) and the Brachycera (short-horned flies) (Fig 1.1). The nematocerans are a presumably paraphyletic assemblage encompassing midges, mosquitos and crane flies. These flies are characterized by the retained primitive features of long antennae and larval mandibles which articulate from side to side, closing against each other. The second major suborder, the Brachycera, appear to have arisen from the nematoceran group Psychodomorpha (Woodley 1989a; Wood 1991; Sinclair 1992; Michelson 1996) or from a combination of Psychodomorpha and Tipulomorpha (Oosterbroek P. 1995). Within the Psychodomorpha, Anisopopidae has been suggested to be the sister group of the Brachycera (Woodley 1989a; Oosterbroek P. 1995; Krivosheina 1998).

Figure 1.1. Overview of major dipteran clades and species representation

Approximate species representation appears in the right column. Bold internodes show robust support for a taxonomic grouping. Parallel branches indicate possible or likely paraphyly. Underscored group names belong indicate nematoceran infraorders. Reproduced with permission of ANNUAL REVIEWS, from Yeates and Wiegmann (1999) in the format Journal via Copyright Clearance Center.



The Brachycera is well established as a monophyletic suborder based on both molecular data and morphological features. These stout bodied flies are characterized by modifications to the larval head capsule and shortened antennae with the apical segments forming a thread-like arista (Woodley 1989a). According to (Yeates 1999), the brachyceran flies can be divided into 4 monophyletic infraorders: Tabanomorpha, Xylophagomorpha, Stratiomyopmorpha, and Muscomorpha. Relationships among these infraorders are currently unresolved (Hennig 1973; Krivosheina 1989; Woodley 1989a; Krivosheina 1991; Griffiths 1994; Sinclair 1994; Nagatomi 1996).

The members of the muscomorph clade are not currently well defined, with only Cyclorrapha and Empidoidea (collectively Eremoneura) firmly established (Chvála 1983; Woodley 1989a; Sinclair 1992; Wiegmann 1993; Griffiths 1994; Cumming 1995). Other possible members of Muscomorpha include Nemestrinoidea (tangle-veined flies and larval parasites of spiders), and Asiloidea (robber flies, stiletto flies, and bee flies) (Woodley 1989a) (Fig 1.1). However these two groups have also been placed in a clade with Tabanomorpha (horse flies) and Xylophagomorpha to form an Asilotabaniform grouping (Griffiths 1994; Zatwarnicki 1996). Muscomorpha is an extremely successful group, encompassing nearly 65,000 flies at its most exclusive (excluding all but the firmly entrenched eremoneurans) to approximately 77,000 species at its most expansive.

A major lineage within Muscomorpha is the Cyclorrhapha (Fig 1.1). Cyclorrhaphan flies possess several distinct larval features which make them easily distinguishable from other flies. The cuticle of the last larval instar of this lineage serves as the puparium. The head capsule of the larva is completely internalized into the thorax, thus the Cyclorrhaphan larva are described as acephalic. Larval mouthparts are also

altered and reduced, with simple hook-like mandibles serving as their sole external feeding apparatus (Griffiths 1972; McAlpine 1981; Stoffolano 1988; Cumming 1995). These alterations to the larval body plan allow the larva to live within its food source, dissolving its environment with saliva and scooping the liquefied food into its mouth.

The Cyclorrhapha can be divided into two groups, a likely paraphyletic group of basal cyclorraphans known collectively as “Lower Cyclorrhapha” or “Aschiza”, and a monophyletic group known as the Schizophora (Griffiths 1972; Griffiths 1991; Wada 1991; Cumming 1995; Zatwarnicki 1996) (Fig 1.1). Lower Cyclorrhapha consists of a handful of small families of flies with the diverse Phoridae (“scuttleflies”) and Syrphidae (“flower flies” or “hover flies”) making up the majority of recorded species (~6000 species in each group) (McAlpine 1981). Relationships within the Lower Cyclorrhapha are disputed, with the small group Opetia widely regarded as the most basal lineage (Griffiths 1972; Griffiths 1991; Wiegmann 1993).

The other major branch of cyclorraphan flies, the Schizophora, account for a large percentage of dipteran diversity, with ~44,000 described species (McAlpine 1981). The Schizophora are united primarily based on the presence of an inflatable head sac called the ptilinum which is used by the adult fly to emerge from the puparium (McAlpine 1981). These flies fall into two groups: the monophyletic Calyptratae, characterized by the presence of well developed calypter at the base of the wing, and the likely paraphyletic acalyptrate flies (Griffiths 1972). There are multiple competing hypotheses regarding classification of these groups with the three most prominent being those put forth by Hennig, McAlpine, and Griffiths (Hennig 1958; Hennig 1971; Griffiths 1972; Hennig 1973; McAlpine 1981). Both McAlpine and Griffiths based their classifications

on the original work by Hennig, with McAlpine refining it and Griffiths proposed a more radical restructuring. In Griffiths' revision he placed all of Schizophora within five superfamilies Lonchaeoidea, Lauxanioidea, Drosophiloidea, Nothyboidea, and Muscoidea (Griffiths 1972). The Muscoidea superfamily contained all of the Calyptratae and many acalyptrate clades, asserting a paraphyletic origin for the acalyptrates. Griffiths made no attempt to resolve relationships between these 5 superfamilies. McAlpine, on the other hand, mostly maintained Hennig's groupings and divided all of Schizophora into 13 superfamilies: the 10 acalyptrate superfamilies Neroidea, Conopioidea, Lauxanioidea, Sciomyzoidea, Ephydroidea, Opomyzoidea, Carnoidea, Sphaeroceroidea, Diopsoidea, and Tephritoidea; and the 3 calyptrate superfamilies Hippoboscoidea, Muscoidea, and Oestroidea (McAlpine 1981). McAlpine attempted to resolve relationships between these 13 superfamilies and arrived at monophyletic Acalyptratae and Calyptratae clades.

The acalyptrate flies are extremely species diverse, with nearly half of dipteran family level diversity belonging to the group (McAlpine 1981). Relationships between these groups are heavily debated with weak support for many theorized clades (Yeates 1999). This is likely due both to a narrower family definition among the acalyptrates than is seen among other fly groups (Yeates 1999), and to the rapid radiation of the cyclorrhaphan clade leading to short internodes, thus leaving few strong synapomorphies to unite them.

Calyptratae is well supported as a monophyletic clade containing the families Calliphoridae, Sarcophagidae, Tachinidae, Anthomyiidae, Muscidae, Streblidae, Nycteribiidae, Hippoboscidae, Glossinidae, and Oestridae (Hennig 1971; Griffiths 1972;

McAlpine 1981). The morphology based classifications of McAlpine and Griffiths agree to a significant degree, differing mostly in naming convention. In McAlpine's phylogeny, Glossinidae, Hippoboscidae, Streblidae, and Nycteribiidae belong to the superfamily Hippoboscoidea, while in Griffiths' schema, this clade is called the Hippoboscidae family grouping (Griffiths 1972; McAlpine 1981). McAlpine and Griffiths also agree on a clade containing Calliphoridae, Mystacinobiidae, Sarcophagidae, Rhinophoridae, Tachinidae, and Oestridae, known as the Oestroidea in McAlpine's classification and the Tachinidae family grouping in Griffiths' work. The two authors disagree on the remaining groups, however. McAlpine places Scatophagidae, Anthomyiidae, Faniidae, and Muscidae into a monophyletic Muscoidea superfamily, while Griffiths considers these groups to be paraphyletic within his Calyptratae prefamily.

CHAPTER 2 "SHAKING THE FLY TREE OF LIFE: PERFORMANCE ANALYSIS OF NUCLEAR AND MITOCHONDRIAL SEQUENCE DATA PARTITIONS"

Introduction

There is a long history of discussion over the phylogenetic utility of mitochondrial versus nuclear genes (Brower, Desalle 1994; Simon et al. 1994; Lin, Danforth 2004; Rubinoff, Holland 2005; Zink, Barrowclough 2008). While it is generally accepted that nuclear genes tend to outperform mitochondrial genes in phylogeny reconstruction on a per site basis (Baker, Wilkinson, DeSalle 2001; Springer et al. 2001; Leys, Cooper, Schwarz 2002; Lin, Danforth 2004; Galewski et al. 2006), these studies have typically focused on the information content of single or small numbers of mitochondrial genes. As one of the properties of likelihood based approaches is consistency (as the amount of data increases towards infinity, the probability of recovering the true tree approaches 1.0) (Fisher 1922), the actual value of utilizing a larger number of mitochondrial sites, such as a full mitochondrial genome, is not clear.

From a data acquisition perspective, mitochondrial gene sequences are more easily obtained due to their high copy number, commonly available conserved primer sets (Simon et al. 1994), lack of introns, and very rare incidence of gene duplication. They are, however, known to evolve rapidly (Brown, George, Wilson 1979), prone to biases in base frequency (Gibson et al. 2005), subject to strand influenced inversions of base composition (Hassanin, Leger, Deutsch 2005; Hassanin 2006), and inherited as a single linkage group (Birky 2001). These attributes typically have a negative impact on phylogenetic tree inference, especially for more ancient divergences (Reed, Sperling 1999; Caterino et al. 2001) (See (Rubinoff, Holland 2005) for review).

Despite these potential shortcomings of the mitochondrial phylogenies, the value of the mitochondrial genome as an independent estimator of animal phylogeny is indisputable (Bourlat et al. 2006; Cameron, Barker, Whiting 2006; Webster et al. 2006; Bourlat et al. 2008; Rota-Stabelli et al. 2010). Nodes where nuclear gene based phylogenies agree with mitochondrial gene derived ones can be considered particularly well supported and independently verified. Many researchers have taken advantage of mitochondrial gene availability to augment nuclear gene data sets. It is notable that mitochondrial gene data has figured prominently in many of the recent Assembling the Tree of Life (AToL) projects (Daly et al. 2010; Jacobsen, Friedman, Omland 2010; Silberfeld et al. 2010). These data sets, with their dense taxon sampling, relatively large gene coverage, and typically robustly supported published topologies present interesting test cases for the phylogenetic utility of mitochondrial gene sequences. In this study, we concentrate on the data set generated by the AToL Diptera project (FLYTREE) (Wiegmann et al. 2011).

These developments notwithstanding, the question remains exactly what is the benefit of mitochondrial data over or in addition to nuclear sequence data. Simulation studies investigate the phylogenetic information content of a parameterized sequence source (Huelsenbeck, Bull, Cunningham 1996; Yang 1998; Conant, Lewis 2001; Jermiin et al. 2004; Townsend 2007). This approach is particularly useful for estimating the sequence sample size necessary to resolve specific nodes (Fischer, Steel 2009). A downside of simulation studies is the narrowing but still existing gap between the behavior of simulated and actual sequences. Further, since animal mitochondrial genomes

have a maximal capacity of less than 20,000 base pairs, there is little incentive to explore the potential of larger sequence sample sizes. We therefore chose to explore empirical data sets to obtain deeper insights into the relative performance of nuclear and mitochondrial genes. Specifically, we used the 42 heavily sequenced Tier 1 taxa of FLYTREE as a test data set for comparing clade recovery with nuclear and mitochondrial genes. These 42 taxa were further refined to produce a 25 taxon data matrix (24 Diptera + one outgroup) with maximum sequence coverage and dense sampling within the higher flies (Brachycera). The nuclear and mitochondrial gene components of this data set were analyzed both together and separately under a variety of partitioning schemes.

We find that within our dipteran test data set, mitochondrial genes, while generally inferior to nuclear genes when analyzed alone, are capable of resolving some relationships for which nuclear genes fail. Moreover, the combined analysis of mitochondrial and nuclear gene produced superior results to either data source alone. In cases where mitochondrial and nuclear gene data sets generated conflicting topologies, the combined data set typically resolved the conflict and produced a topology consistent with current hypotheses with no loss of branch support. Our results also yield important insights regarding the robustness of previously inferred topologies in the phylogeny of Diptera.

Methods

Sequence alignment

Single gene, codon consistent nucleotide sequence alignments were produced with MEGA 4.0 (Kumar et al. 2008) . Variable sites and regions of poor alignment were removed using Gblocks (Talavera, Castresana 2007) in codon mode with default block

parameters and a 50% missing sites threshold. In addition to the mitochondrial and nuclear gene alignments, a concatenated alignment was created. All trimmed alignments have been deposited as supplementary data.

Bayesian tree construction

Tree reconstruction was performed on the Wayne State University High Performance Computing Grid. Bayesian trees were constructed using MrBayes v3.1.2 compiled for MPI systems (Huelsenbeck, Ronquist 2001; Ronquist, Huelsenbeck 2003; Altekar et al. 2004). For all data sets, two independent runs of four chains were run for five million generations with sampling every 100 generations and 25% of samples discarded as burn-in. Each data partition was assigned an independent model with a gamma rate heterogeneity parameter and an invariable sites parameter. For nucleotide data sets, each partition was assigned a GTR model. Convergence was checked for each data set after sampling was completed.

Tree analysis

Custom Perl scripts (available upon request) using Bioperl (Stajich et al. 2002) and Bio::Phylo (Vos et al. 2011) were written to parse tree data and generate summaries.

Results

Data matrix preparation

Taxa for our analyses were selected from the Tier 1 species of the FLYTREE project (Wiegmann et al. 2011), which give a balanced sampling of dipteran diversity and provide broad coverage of important divergences. As anchor points for the backbone dipteran phylogeny, the Tier 1 taxa have been sequenced for their entire mitochondrial genome and 12 single copy nuclear protein coding genes. In contrast, only 5 nuclear

genes have been sequenced for the Tier 2 taxa (Wiegmann et al. 2011). While recent work suggests that the incorporation of incompletely sampled genes may have some beneficial effects on tree reconstruction (Burleigh, Hilu, Soltis 2009), this represented a special case where incompletely sampled genes were added to a complete data set. Moreover, previous work suggests an overall negative effect of missing data on phylogenetic inference for irregularly distributed missing data (Wiens 1998; Hartmann, Vision 2008). Thus taxa for which less than 75% of the total sequence length was present were discarded to minimize the potential negative effects of gaps. For these, all thirteen protein coding genes from the mitochondrial genome were concatenated and 12 protein encoding nuclear genes were selected for analysis. Subsequent application of the Gblocks program (Castresana 2000) further reduced the amount of missing data by removing sites which were present for fewer than 50% of the included taxa.

The resulting data matrix contained twenty four Diptera and one outgroup (*Tribolium castaneum*) with mitochondrial and nuclear genes extensively sampled (Fig. 1). As taxon sampling in the non-brachyceran flies was uneven and preliminary investigations showed a great deal of instability in this part of the tree for both mitochondrial and nuclear encoded genes (not shown), only a single representative of Culicomorpha and Tipulomorpha and two representatives of Bibionomorpha were retained. Four species representing most major lineages of the basal "orthorrhaphous" Brachycera (Tabanamorpha, Stratiomyomorpha, and two representatives of Asiloidea) were included, as was a specimen from Empididae, a basal member of the Eremoneura clade. Within the Cyclorrhapha, three "lower" cyclorrhaphans (Phoridae, Lonchopteridae, and Syrphidae) were included. Five non-calyptrate schizophorans

(Drosophilidae, Sepsidae, Lauxaniidae, Diopsidae, and Tephritidae) were selected based upon the sequence coverage criteria described above. Lastly, seven representatives of the Calyptratae (Glossinidae, Muscidae, Scatophagidae, Anthomyiidae, Sarcophagidae, Tachinidae, and Calliphoridae) were selected to provide a comprehensive sampling of major families.

The mitochondrial alignment included 10,812 base pairs after removal of variable and poorly represented sites, which compared with 6,528 nucleotide sites in the nuclear alignment. The concatenated sequence of mitochondrial and nuclear genes contained 17,340 base pairs.

Establishing benchmark clades

In order to avoid the circular condition of assessing clade robustness based on our own consensus results, only clades consistently recovered in both Wiegmann *et al.*(2011) and in our analyses were considered as potential benchmark clades (Fig 2.1). Clade support was classified in 3 categories (Table 2.1).

"Robust" status indicated consistent support for a clade with no competing signal. "Robust" clades were recovered by at least one concatenated (mitochondrial and nuclear) gene data set. Moreover, "robust" clades were also recovered by at least mitochondrial or nuclear genes alone, although not necessarily by both sets. Lastly, these clades were recovered by more than one codon position or codon position data set combination.

Reassuringly, the vast majority of clades were recovered with robust support across multiple data sets (Table 1) and included all well established monophyletic groups (Brachycera, Eremoneura, Cyclorrhappha, Schizophora, and Calyptratae) (Fig 2.1, nodes 4, 8, 9, 12, 14), although mitochondrial genes alone failed to resolve Eremoneura and

Figure 2.1. Dipteran phylogeny

Tree topology arrived at by Wiegmann (2011). Numbers at nodes indicate identifier number for clade. Clade ages derived from Wiegmann *et al.* (2011) and Grimaldi and Engel (2005).

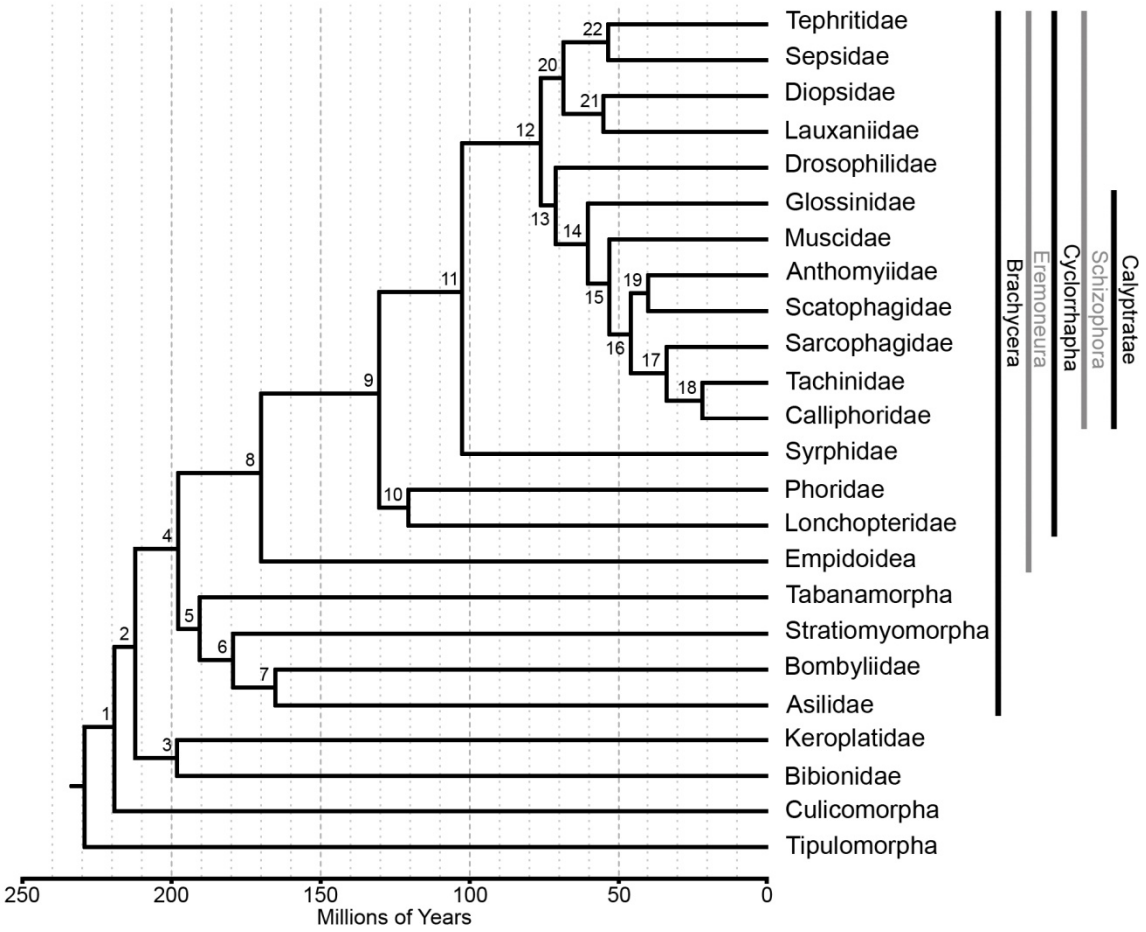


Table 2.1. Clade support by data partition

Node numbering is according to Fig 2.1. "-" indicates that the clade was not recovered by the dataset. Clades in bold were clades included in the high confidence data set. Clades in italics are clades which fell into the moderate support category. Green = posterior probability > .80. Yellow = posterior probability =<.80. Alternative topologies are clades which we tested which do not match those of Fig 2.1. Muscomorpha: Asiloidea+ Eremoneura. Brach-Tab: Basal position of Tabanamorpha relative to the remaining Brachycera. Sarc+Call: Sarcophagidae + Calliphoridae. Diop + Teph: Diopsidae + Tephritidae. Mit Oest + Musc1 & Musc 2: Recovery of clades 15 and 16 corrected for erroneous placement of Tachinidae in mitochondrial data sets.

		Nuclear					Mitochondrial					Nuclear + Mitochondrial				
node	Type	1st	2nd	3rd	12	all	1st	2nd	3rd	12	all	1st	2nd	3rd	12	all
1	Culico+Neo	-	-	-	-	-	-	0.76	-	-	-	-	0.91	-	-	-
2	***Neodiptera***	-	***0.86***	-	***0.97***	***1.00***	-	-	-	-	-	***0.84***	-	-	***0.65***	-
3	**Bibionomorpha**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.77**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	**Brachycera**	**1.00**	**1.00**	-	**1.00**	**1.00**	-	**1.00**	-	**0.64**	-	**1.00**	**1.00**	-	**1.00**	**1.00**
5	Orthorrhapha	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6		-	-	-	**0.54**	-	-	**0.96**	-	**0.64**	-	**1.00**	**1.00**	-	**1.00**	**1.00**
7	**Asiloidea**	**0.99**	**0.95**	-	**1.00**	**1.00**	-	-	-	-	-	**1.00**	**0.58**	-	**1.00**	**1.00**
8	**Eremoneura**	**0.88**	**0.99**	-	**1.00**	**0.60**	-	-	-	-	-	**1.00**	**0.99**	-	**1.00**	**1.00**
9	**Cyclorappha**	**1.00**	**0.99**	-	**1.00**	**1.00**	**1.00**	-	-	**1.00**	**1.00**	**1.00**	**1.00**	-	**1.00**	**1.00**
10	**Platypezoidea**	-	-	-	-	-	**1.00**	-	-	**1.00**	**1.00**	**0.96**	**0.98**	-	**1.00**	**1.00**
11		-	**1.00**	-	**1.00**	-	**1.00**	**1.00**	-	**1.00**	**1.00**	**1.00**	**1.00**	-	**1.00**	**1.00**
12	**Schizophora**	**1.00**	**1.00**	-	**1.00**	**1.00**	-	**0.80**	-	-	-	**1.00**	**1.00**	-	**1.00**	**1.00**
13		-	**0.96**	-	-	-	**0.99**	-	-	**1.00**	**1.00**	-	**0.87**	-	**0.87**	-
14	**Calyptratae**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
15		**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	-	**1.00**	**1.00**	**1.00**	**1.00**	**0.96**	**1.00**	**1.00**
16		**1.00**	**1.00**	-	**1.00**	**1.00**	-	-	-	-	-	**1.00**	**1.00**	**0.98**	**1.00**	**1.00**
17	**Oestroidea**	**1.00**	**0.98**	-	**1.00**	**1.00**	-	-	-	-	-	**1.00**	**0.90**	-	**1.00**	**1.00**
18	Tach+Call	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	**Anth+Scat**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
20		-	***0.73***	-	-	-	-	-	-	-	***0.54***	-	***0.51***	-	***0.83***	-
21	Sep+Teph	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	Diop+Laux	-	-	-	-	0.75	-	-	-	-	-	-	-	0.89	-	-
		Alternative topologies														
15*	Mit Oest+Musc2	-	-	-	-	-	1.00	-	1.00	1.00	1.00	-	-	-	-	-
16*	Mit Oest+Musc1	-	-	-	-	-	1.00	-	0.97	1.00	1.00	-	-	0.81	-	-
23*	Muscomorpha	-	0.65	-	-	-	-	-	-	-	-	-	-	-	-	-
24*	Brach-Tab	1.00	-	-	1.00	1.00	-	-	-	0.64	-	1.00	1.00	-	1.00	1.00
25*	Sarc+Call	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
26*	Diop+Teph	0.95	0.99	-	1.00	-	-	1.00	-	1.00	1.00	0.98	1.00	-	1.00	1.00

performed poorly in recovering Schizophora. The clade which joins Asiloidea and Stratiomyomorpha (Fig 2.1, node 6) was a borderline case for inclusion into the "robust" category, being well resolved only by concatenated data sets and mitochondrial 2nd codon positions. However posterior probabilities were very high for this clade and no competing topologies were consistently recovered by other data sets.

"Moderate" support for a clade was assigned if there was a high degree of sensitivity to codon position inclusion, consistent recovery by only a single data source set (mitochondrial, nuclear, or concatenated), or generally low support values (<=.80 posterior probability). This classification represented clades which were inconsistently recovered but for which the consensus of evidence was supportive and no strong competing signal was indicated. Only two clades fell into the "moderate" category. The Neodiptera were poorly supported in several analyses (Fig 2.1, node 2), with no support from mitochondrial data and only very weak support from concatenated data sets. Second, the clade containing all non-*Drosophila* "acalyptrate" flies (Sepsidae, Lauxaniidae, Diopsidae, and Tephritidae) (Fig 2.1, node 20) was recovered by four data sets, however support values for this clade ranged from only .51 to .83 (Table 2.1, node 20).

Lastly, we classified clades as having "low" support if the results indicated the presence of a strong signal for a competing topology or very little support for any given topology. Clades were also assigned to the "low" support category if our results consistently recovered a topology which differed from the topology recovered by Wiegmann *et al.* (2011). The latter condition was encountered for 5 FLYTREE clades. The clade Culicomorpha + Neodiptera (all flies excluding Tipulomorpha) (Fig 2.1, node

1) was recovered by only two data sets with support values of .76 and .91. This was possibly a rooting artifact due to using a distantly related coleopteran as the outgroup. Second, the clade Orthorhappha (Fig 2.1, node 5), newly reintroduced in Wiegmann *et al.* (2011), was not recovered by any of our data sets. Instead, Tabanamorpha was consistently inferred to be the oldest brachyceran group, and sister to all remaining Brachycera. Third, the clade Sepsidae + Tephritidae (Fig 2.1, node 21) was not recovered by any of our trees. Next, the clade Lauxaniidae + Diopsidae (Fig 2.1, node 22) was only recovered in two trees. Instead, Diopsidae + Tephritidae was recovered in multiple trees and the position of Sepsidae was unstable. Finally, the calyptrate clade Calliphoridae + Tachinidae (Fig 2.1, node 18) was never recovered in our trees while an alternative clade Sarcophagidae + Calliphoridae was recovered by every data set.

Performance of mitochondrial, nuclear, and concatenated data sets

The 17 robust and moderately supported benchmark clades afforded us the opportunity to systematically compare how well mitochondrial and nuclear data sources performed on their own and in combination. Bayesian trees were estimated for the single trimmed mitochondrial and nuclear alignments as well as for the concatenated alignments. These three basic approaches were performed for combined as well as separate codon positions, resulting in a total of 15 trees and 255 branches for comparative analysis (Fig. 1).

Across all trees considered, we found that mitochondrial and nuclear genes performed comparably in their ability to resolve clades. At least one of the nuclear gene codon position sets was able to recover 16 of the 17 high confidence benchmark nodes. Mitochondrial genes alone recovered only 12 of those nodes, however two nodes were

lost due to an obviously erroneous placement of the tachinid fly *E. larvarum* in a position basal to the Muscomorpha grade (Fig 2.1, between nodes 14 and 15)(Kutty et al. 2008). If dispensation was made for this (Table 2.1, nodes 15*, 16*), mitochondrial gene clade recovery rose to 14 nodes.

Most significantly, both mitochondrial and nuclear genes were capable of recovering clades, which the other data set was not. Mitochondrial genes recovered the Platypezoidea clade (Fig 2.1, node 10) while nuclear encoded genes did not. Nuclear genes, on the other hand, could recover monophyletic Neodiptera, Asiloidea, Eremoneura, and the proper position of the tachinid *E. larvarum* within the Calyptratae (Fig 2.1, nodes 2, 7, 8, 15). As neither data set was capable of recovering the complete set of 17 nodes on its own, the value of combining mitochondrial and nuclear genes in tree estimation was readily apparent.

Relative performance of first and second codon positions

Since mitochondrial and nuclear genes recovered select clades which the other data set did not, we examined if this discordance could be mitigated by more specific codon partition choices. In the nuclear gene data set, 2^{nd} codon positions alone greatly outperformed 1^{st} codon positions. The former were capable of resolving 15 of the 17 benchmark clades while the latter resolved only 11 (Table 2.1). In contrast, the mitochondrial 1^{st} or 2^{nd} codon position data sets performed comparably to each other with each recovering 8 benchmark clades (Table 2.1).

Strikingly, we found several cases where single codon positions (1^{st} or 2^{nd} codon positions alone) recovered nodes that the more inclusive nuclear or mitochondrial data sets (1^{st} + 2^{nd} or 1^{st} + 2^{nd} + 3^{rd} codon position) did not. In mitochondrial gene data sets,

for instance, the Schizophora clade (Fig 2.1, node 12) was recovered only by 2nd codon positions alone. With the nuclear encoded genes, 1st + 2nd codon positions failed to resolve the sister group relationship between Drosophilidae and Calyptratae (Fig 2.1, node 13) and did not group the four remaining non-calyptrate schizophorans into a monophyletic clade (Fig 2.1, node 20) while 2nd codon positions were capable of recovering these relationships. This indicated that the evolutionary pattern or rates differed between these two codon positions.

Furthermore, there were only two cases of well supported nodes where the inclusion of more than one codon position in the data set was necessary for recovery of a node that single codon positions failed to recover. In one of these cases, the nuclear gene 1st and 2nd codon positions combined recovered the monophyletic clade containing the Asiloidea and Stratiomyomorpha (Fig 2.1, node 6) with low support values, but no single codon position from the nuclear genes could resolve this clade on its own (Table 2.1, node 6). In the second case, the mitochondrial genes recovered the monophyletic clade containing Tephritidae, Sepsidae, Lauxaniidae, and Diopsidae (Fig 2.1, node 20) when all three codon positions were included, but no single codon position alone recovered the clade (Table 2.1, node 20).

As single codon positions proved to have phylogenetic utility similar to the more inclusive 1st + 2nd codon position data sets, we finally examined the congruence between clades recovered in the separate analyses of 1st or 2nd codon positions . In the mitochondrial gene trees, we found surprisingly little overlap between clades recovered by 1st codon positions and clades recovered by 2nd codon positions. There were only four clades, which were recovered by both mitochondrial 1st codons and mitochondrial 2nd

codons (Table 2.1, nodes 3, 11, 15, 19). Three clades were only recovered by 2nd codon positions but not 1st (Table 2.1, nodes 4, 6, 12) and four clades were recovered by 1st but not by 2nd codon positions (Table 2.1, nodes 9, 13, 15*, 16*). Importantly, all of these clades except for two (12 and 13) were recovered by combined mitochondrial 1st and 2nd codon position data sets. Thus, the poor clade recovery of single codon positions alone may be merely the consequence of insufficient sequence length in the individual codon position data sets rather than conflicting or misleading signals between codon site partitions.

Trees generated from single codon positions in the nuclear gene data set showed a much more consistent distribution of phylogenetic signal. In all cases where only one of the 1st codon or 2nd codon position data sets recovered a clade, it was always the 2nd codon positions that recovered the clade. Most high confidence clades which were recovered by either 1st or 2nd codon positions alone were recovered by both data sets.

Finally, we discovered that using concatenated mitochondrial and nuclear genes, all high confidence clades were recovered by either 1st or 2nd codon positions alone. Second, the majority were recovered in both 1st and 2nd codon position data sets. Taken together, the codon specific analyses underlined the improvement of robust tree estimation performance gained by combining mitochondrial and nuclear sequence data and suggested that the phylogenetic signal of mitochondrial gene data is evenly split between 1st and 2nd codon positions.

Performance of third codon positions

Rapid accumulation of substitutions at 3rd codon positions is known to lead to saturation at those sites and degradation of phylogenetic signal. Removal of 3rd codon

positions from a protein coding data set is therefore a standard procedure in phylogenetic inference. In our analyses, both mitochondrial and nuclear 3rd codon positions showed approximately equal phylogenetic utility, but it was extremely low. Interestingly, however, 3rd codon positions were capable of resolving some recent clades within the Calyptratae (Fig 2.1, nodes 14, 15, 16, 19) and the monophyly of the two bibionomorph taxa (Fig 2.1, node 3).

When 3rd codon positions were combined with 1st and 2nd, their negative impact on tree reconstruction was minor. Within the mitochondrial gene results, the clade Brachycera (Fig 2.1, node 4) and the sister group relationship between Asiloidea and Stratiomyomorpha (Fig 2.1, node 6) was recovered by 1st + 2nd codon position data sets but not 1st + 2nd + 3rd. Similarly, nuclear genes trees failed to recover the Asiloidea + Stratiomyomorpha clade (Fig 2.1, node 6) and the sister group relationship of Syrphoidea to Schizophora (Fig 2.1, node 11) when 3rd codons were included. When nuclear genes were concatenated with mitochondrial genes, the Neodiptera clade (Fig 2.1, node 2) and internal relationships within the non-calyptrate schizophorans (Fig 2.1, nodes 13, 20) were recovered with 1st + 2nd but not 1st + 2nd + 3rd. In one case, Tephritidae + Sepsidae + Lauxaniidae + Diopsidae (Fig 2.1, node 20), the 1st + 2nd + 3rd mitochondrial gene data set was able to recover a node that was not resolved by 1st + 2nd alone, however this was the only case where 3rd codon inclusion apparently improved clade recovery. Taken together, these results lent further support to the practice of excluding 3rd codon positions, if only for the effect of reducing computational burden.

Discussion

Mitochondrial sequences are highly beneficial in large scale tree reconstruction

Our data set allowed the analysis of both mitochondrial and nuclear gene sources as independent estimators of phylogenetic relatedness. While the utility of the mitochondrial genome in resolving some deep level dipteran relationships has been already shown (Cameron et al. 2007), the comparison of relative phylogenetic utility between mitochondrial and nuclear data sources remains a topic of interest.

As demonstrated by our results, full length mitochondrial genome data sets possess sufficient phylogenetic signal to resolve nearly all nodes we tested in the dipteran phylogeny. As this group's history spans a large time depth, with nodes ranging from approximately 30-250 million years divergence time and contains several major radiations characterized by very short internodes, this real world data set represents a non-trivial test case for data performance. Further, while we have found that nuclear genes display more consistent behavior than mitochondrial genes, we observed superior clade recovery when both mitochondrial and nuclear genome data are included in the same analysis. Importantly, our finding that mitochondrial gene data proved superior in resolving some nodes which the nuclear gene data performed poorly on suggests that the synergistic effect of the combined analysis was not simply due to the sequence sample size increase. It seems reasonable to predict that the concatenation of mitochondrial and nuclear gene sequences generally provides results that cannot be obtained from small data sets containing nuclear genes alone. Taking further into account the relative ease of mitochondrial genome acquisition and the lack of any obvious deleterious effects on tree

reconstruction in combined analysis, mitochondrial gene data inclusion is undeniably effort and cost efficient in increasing overall tree robustness.

From a data analysis perspective, we have also shown that nuclear genes display more consistent behavior than mitochondrial genes; however several nodes were not adequately resolved by nuclear genes alone. As such, we conclude that concatenation of mitochondrial and nuclear gene sequences provides superior results that can not be obtained from small data sets containing nuclear genes alone. While broad phylogenetic questions have become a matter of genome-wide phylogenetic analyses with the advent of next generation sequencing technologies, the design of sequencing strategies for the comprehensive phylogenetic analysis of extremely species-rich clades such as the Diptera (Baker, Wilkinson, DeSalle 2001; Cameron et al. 2007; Dyer et al. 2008; Gibson, Skevington, Kelso 2010; Singh, Kurahashi, Wells 2011) will continue to depend on the herein confirmed benefit of mitochondrial genomes for time to come.

Brittle branches in the fly tree of life

The bursts of explosive radiations that characterize the megadiverse Diptera (Wiegmann et al. 2003; Wiegmann et al. 2011) make establishing a robust phylogeny a challenging endeavor. It has been shown that the amount of homologous sequence data may be more important than taxon sampling in phylogeny reconstruction (Rokas, Carroll 2005). The comparison of the topology obtained from combined analysis with the more completely sequenced 25 taxon data set we constructed with the conclusions in Wiegmann *et al.* (2011) is therefore a useful test of dipteran clade robustness.

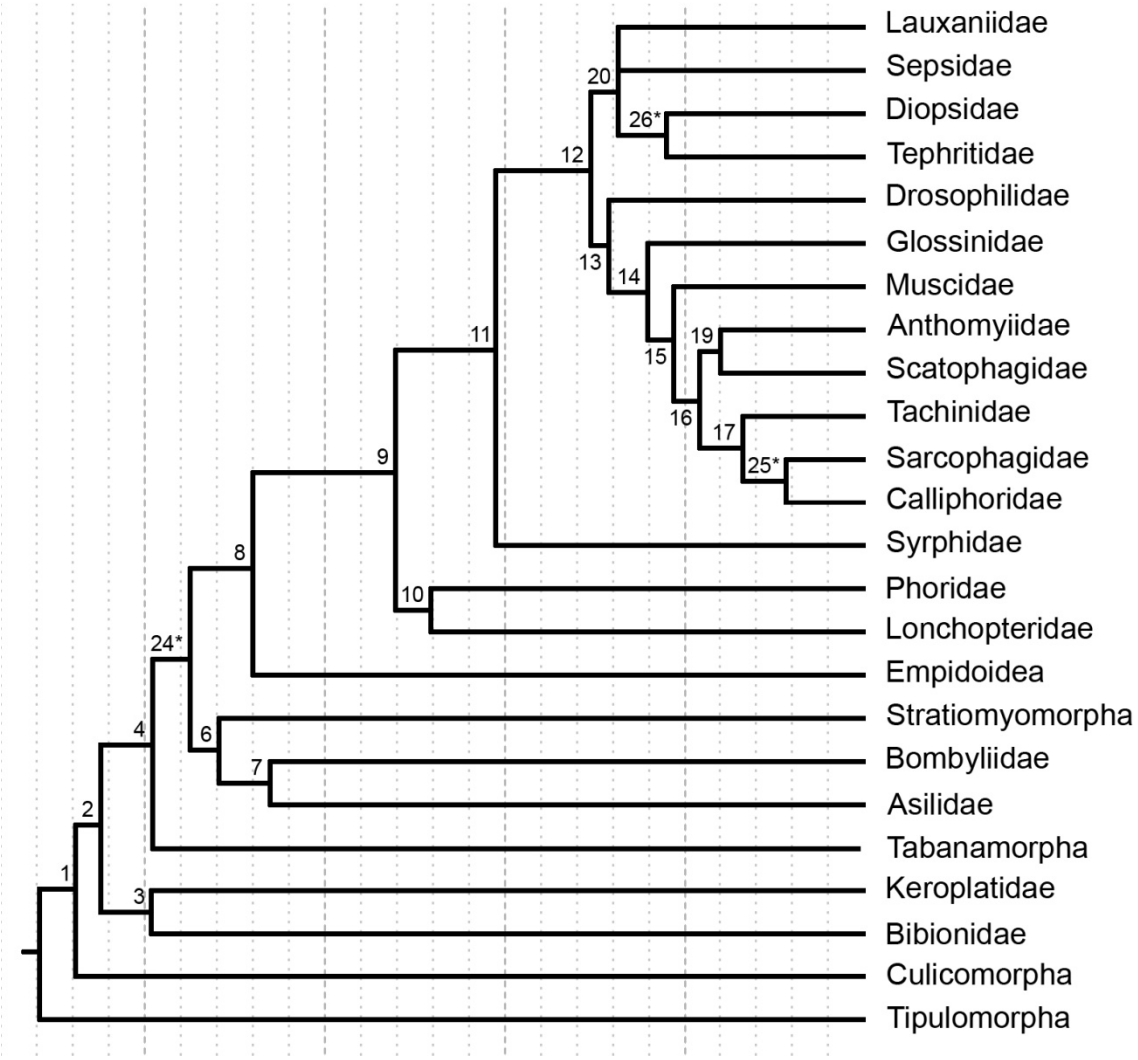
Our data sets were derived from those of Wiegmann *et al.* (2011), but differed dramatically in taxon sampling, composition, and site coverage. While

comprehensive/exhaustive taxon sampling was the goal in Wiegmann *et al.* (2011), our data set emphasized maximum sequence coverage and, more importantly, consistent inclusion of mitochondrial gene data as well as nuclear gene data. Gratifyingly, our analyses produced results largely congruent with those of Wiegmann *et al.* (2011). All historically well supported monophyletic clades (Brachycera, Eremoneura, Cyclorrhapha, Schizophora, Calyptratae) (Fig 2.2 and 3, nodes 4, 8, 9, 12, 14) were robustly recovered. Moreover, Neodiptera (Fig 2.2 and 3, node 2) was confirmed with moderate support and Bibionomorpha was corroborated as the sister group to Brachycera (Fig 2.2 and 3, nodes 2 and 4). Stratiomyomorpha was recovered as the sister group of Asiloidea (Fig 2.2 and 3, nodes 6 and 7). Finally, Drosophilidae, representing the Ephydroidea, was often recovered as the sister group to Calyptratae (Fig 2.2 and 3, nodes 13 and 14), although in some cases a Drosophilidae + Sepsidae clade was supported as the sister to Calyptratae.

However, we were unable to confirm some of the more surprising or tentative conclusions of the FLYTREE project (Fig 2.2). The most notable disagreement between our results and those of Wiegmann *et al.* (2011) is in how basal Brachyceran groups were arranged. Our trees failed to recover the monophyletic Orthorrhapha clade (Tabanamorpha + Stratiomyomorpha + Asiloidea) (Fig 3, node 5) supported by Wiegmann *et al.* Our results instead strongly suggest that Tabanamorpha is the most basal brachyceran group, sister to the remaining Brachycera (Fig 2.2, nodes 4 and 24*). Similarly, we failed to recover as monophyletic the Muscomorpha clade (Asiloidea + Eremoneura) (Table 2.1, node 23*), which is one of the more common alternative topologies for the brachyceran infraorders (Woodley 1989b; Yeates, Wiegmann 1999).

Figure 2.2. Consensus topology

Tree topology arrived at by our analyses. Nodes not present in Fig 2.1 are marked with an "*".



Our results instead indicate that a clade containing Stratiomyomorpha and Asiloidea should be placed as the sister group to Eremoneura (Fig 2.2, nodes 24* and 8).

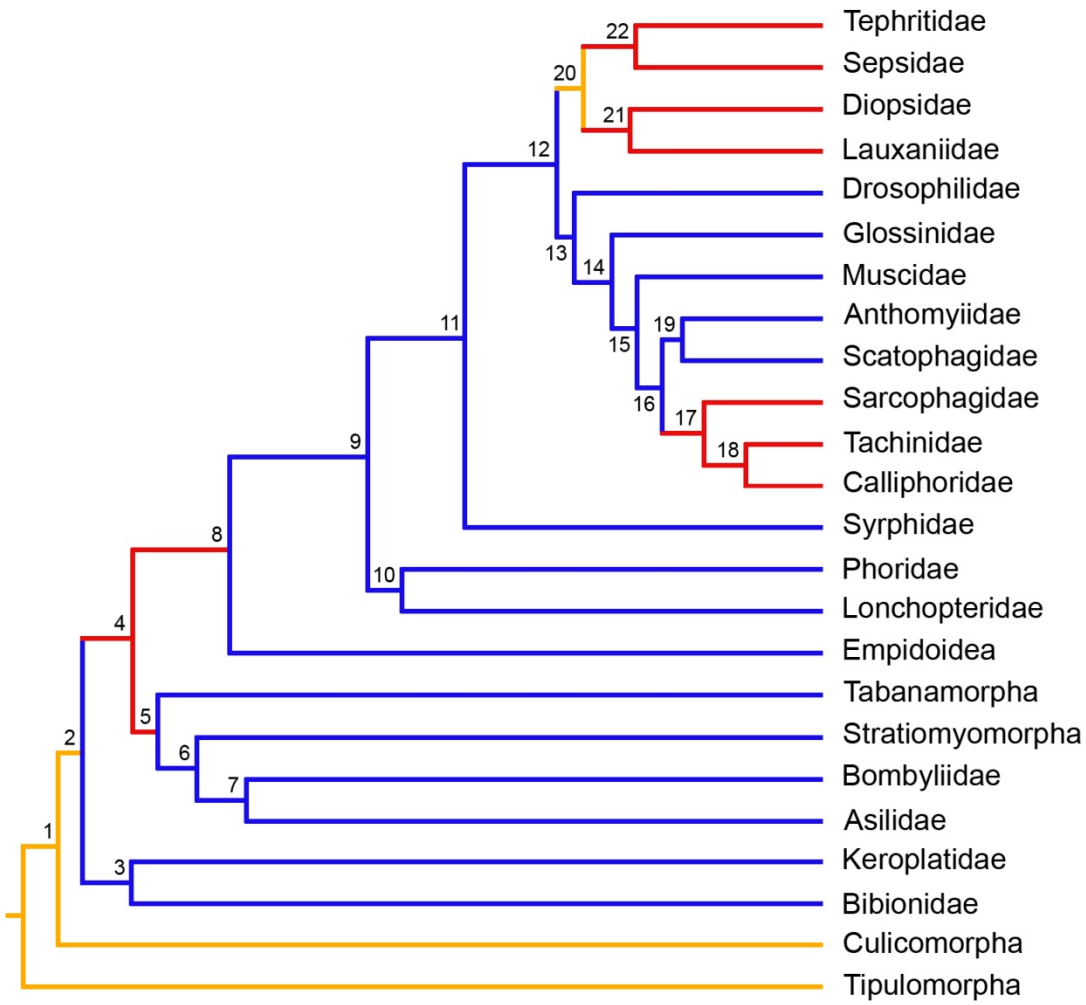
Furthermore, our results suggest that the relationships among acalyptrate flies are far from firmly established. Aside from the placement of a group containing the Drosophilidae as sister to the Calyptratae (Fig 2.2 and 2.3, nodes 13 and 14), there is little agreement in the topology of non-calyptrate fly relationships between our trees and those of Wiegmann *et al*. (2011). As this area of the tree likely suffers from sparse taxon sampling in our analyses, the Wiegmann *et al.* (2011) acalyptrate relationships may be considered more informative. However, many of the Wiegmann *et al.* estimates for these relationships suffer from low branch support. Therefore, we conclude that non-calyptrate fly relationships should be considered tentative at this point, remaining an important challenge for future studies by dipteran phylogeneticists. The methodological results of our study allow for the prediction that expanding the combined mitochondrial and nuclear sequence coverage for the tier 2 level taxon sample will lead to substantial improvements in this and other problematic areas of the fly tree of life.

Acknowledgements

We thank all our collaborators on the fly tree of life project which was funded by NSF award EF-0334948. J.C. was further supported by a Wayne State University Graduate Enhancement Research Fellowship. Computational resources were provided by the Wayne State University High Performance Computing Grid.

Figure 2.3. Robustness of dipteran clades

Branches in blue are robustly supported by our results. Branches in yellow are moderately supported by our results. Branches in red were not recovered or were weakly recovered in our results.



CHAPTER 3 "MITOCHONDRIAL VERSUS NUCLEAR DNA DERIVED DIVERGENCE TIME ESTIMATES: A CASE STUDY IN THE HIGHER DIPTERA"

Introduction

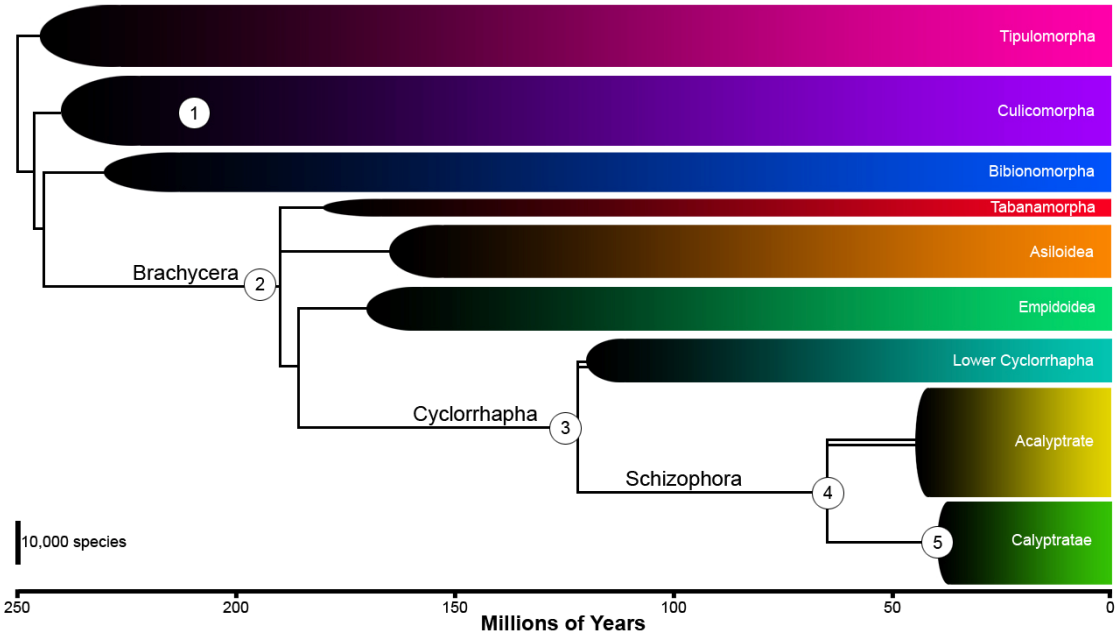
The application of rates of molecular evolution to the determination of species divergence times has a long history (Sarich, Wilson 1967a; Sarich, Wilson 1967b; Wilson, Sarich 1969) but its role in questioning the "Cambrian explosion" (Wray, Levinton, Shapiro 1996) has brought it into a recent vogue. Species divergence time estimates are becoming very common corollary additions to phylogenetic studies, yet the overall accuracy of these estimates has not received a thorough evaluation. Studies comparing algorithms and simulation study data abound (Drummond et al. 2006; Yang, Rannala 2006; Lepage et al. 2007; Svennblad 2008; Wu, Susko, Roger 2008) but comparisons between the ability of independent data sources to date the same nodes are scarce. When confronted with divergence time estimates in a manuscript, many readers are unable to critically evaluate the methods through which the dates that were derived and what biases may be present in the data or methods. Furthermore, when researchers embark on their own divergence time estimations, there is little guidance as to which genes may give the best results, which sites should be included, or over what time scales divergence estimates may be most accurate. We set out to address these questions by comparing divergence time estimates in the Diptera using mitochondrial encoded and nuclear encoded genes as independent estimators of clade age. This opportunity presented itself with the accumulation of a substantial body of mitochondrial and nuclear gene sequences in the course of the collaborative effort to resolve the dipteran tree of life (Wiegmann et al. 2011).

The extant Diptera represent one of the largest animal groups with a well-developed phylogenetic framework and equally well-researched paleontological record (Yeates, Wiegmann 1999). It is estimated that the Diptera first appeared approximately 245 mya in the early Triassic (Fig 3.1). The major basal fly infraorders (comprising the "nematoceran" flies) are considered to appear in the fossil record by the late Triassic (Grimaldi, Engel 2005). An abundant fossil record documents the diversification of brachyceran flies in the time period between 187 and 70 mya (Grimaldi, Engel 2005). More recent fossils, however, are sparse. The dearth of late fossils leaves significant questions about the timing and pace of evolution in one of the most recent and most successful clades of flies, the Schizophora. As roughly one third of the extant flies belong to the Schizophora, there is a significant gap in our understanding of recent evolution in the Diptera.

Currently, both mitochondrial genes and nuclear encoded genes are being used for the estimation of divergences without any apparent preferences beyond data availability. Yet these data sources are known to evolve very differently, even to the point of producing dramatically different trees when used for that purpose (Springer et al. 2001; Zink, Barrowclough 2008; Caravas, Friedrich 2010). Considering the long and lively debate regarding which data source is more suitable to which questions of tree reconstruction, the silence on their applications to dating clades is notable. There is only one study which analyzed mitochondrial data and nuclear data separately for the same group (Yang, Rannala 2006); however, only one node between the two data sets is directly comparable and the clade under study (primates) has no nodes older than 35my present in either tree.

Figure 3.1. Approximate ages and taxonomic representation of major dipteran lineages

Vertical height of each group corresponds to approximate species number. Horizontal scale indicates approximate ages of clades and diversification events. Parallel lines indicate possible paraphyly. Numbers in circles indicate calibration points: 1 = 210 my for Brachycera/Culicomorpha split (*Aenne – Grauvogelia*) (Grimaldi, Engel 2005); 2 = 195 my for Brachycera (*Oligophyrne*) (Grimaldi, Engel 2005); 3 = 125 my for Cyclorrhapha (*Opetiala*) (Grimaldi, Engel 2005); 4 = 64 my for Schizophora (*Phytomyzites*) (Winkler et al. 2010); 5=42 my for the Anthomyiidae/Scatophagidae split (*Protanthomyia*) (Grimaldi, Engel 2005).



Under ideal circumstances, clade age estimates derived from independent data sources such as mitochondrial and nuclear genome encoded genes should produce similar divergence time estimates, as elapsed time since species divergence must remain constant. In reality, however, we see dramatic differences in substitution patterns which are known to have significant effects on phylogenetic tree reconstruction efforts. As divergence time estimation software relies on models and methods, which are very closely related to tree reconstruction methods, it is reasonable to assume that similar issues may be encountered when comparing mitochondrial and nuclear gene derived clade ages. Most methods for determining clade ages have stricter requirements than phylogenetic tree reconstruction methods, such as requiring a fixed tree topology and requirements that branches to be strictly scaled according to an absolute time scale rather than allowing each branch length to fluctuate freely. With some of the flexibility removed from the models, it is not clear what effect choosing mitochondrial or nuclear genes will have on the final node age estimates.

Furthermore, the different modes of inheritance between mitochondrial and nuclear genes may be a factor in their utility as age estimators. It is well known that individual genes may have a different history than the actual species history due to the effects of lineage sorting, introgression, and horizontal gene transfer. It is also accepted that sampling multiple genes that are not genetically linked can overcome the possible biases present in a single gene because independent loci are unlikely to share the same tangled history of inheritance (Pamilo, Nei 1988). By sampling multiple loci, a consensus history can be obtained. This holds true for nuclear genes which are usually distributed across multiple large paired chromosomes that are capable of recombining

and breaking genetic linkages over time. Due to this, any two genes picked at random from the genome are extremely unlikely to be strongly genetically linked to one another. The mitochondrial genome, on the other hand, is inherited as a single linked unit, which rarely recombines. Mitochondrial genes, therefore, can not be viewed as independent from one another and will more likely reflect the same history. Also unlike nuclear genes, where chromosome inheritance from a hybrid is randomized in future generations leading to the breakup and possible loss of introgressing genes, the mitochondrial genome is a single entity that is usually inherited from the mother (Birky 2001). Every offspring of a hybridization event will carry the mitochondrial genome of the mother and it will be passed along the maternal line without change. Therefore, if the taxa under investigation underwent extended periods of hybridization and/or introgression, there is a high likelihood of possible mitochondrial contamination from sister taxa.

Here we present the results from an in depth analysis of nuclear versus mitochondrial sequence based divergence time estimates for a representative sample of dipteran species with specific focus on events in the Brachycera. In side by side comparisons of divergence dates from nuclear and mitochondrial gene data, we compare their effectiveness in resolving divergences over a 200 million year time frame. We further investigate the value of third codon positions, utilization of more complex models of evolution, and the effects of alternate data partitioning schemes on clade age recovery.

Materials and methods

Taxon selection

Taxa were selected to provide high resolution at the family level within the Cyclorrhapha as described for the Tier 1 taxa in (Wiegmann et al. 2011) (Fig 3.1, Table

3.1). Five acalyptrate families, eight calyptrate families, and three non-schizophoran cyclorrhaphan families provide a broad sampling of diversity across the full span of cyclorrhaphan evolution. Additional groups were added to mark significant historical points in the tree, including an empid fly to mark Eremoneura, a tabanamorph and two Asiloidea to mark the Brachycera, and a culicomorph for a nematoceran outgroup.

Sequencing

Individual specimens were ground in the presence of protease K, and total genomic DNA was extracted using a standard phenol–chloroform extraction protocol (Stewart, Beckenbach 2003) and Nucleospin DNA purification columns (Macherey-Nagel). An alignment of dipteran and outgroup mitochondrial genomes was used to identify conserved regions. At conserved coding regions approximately 500 bp apart, degenerate primers were designed against both the J and N strand. Primer pairs spanning approximately 1kb were selected for PCR to create two-fold overlapping coverage. The degenerate primer set typically amplified between 40% and 70% of the total coding material. Primer walking was used to cover regions which the degenerate primer set failed to amplify. PCR fragments were purified using the QIAquick PCR Purification kit (Quiagen) and sequenced using Big Dye Terminator sequencing. Base calling was performed using Phred (Ewing, Green 1998; Ewing et al. 1998) and contig assembly was done using Phrap. Contigs were visualized and manually joined using BioLign v4.0.6 (Tom Hall, NC State Univ.).

Mitochondrial genome sequences for *Cyrtodiopsis dalmanni, Delia radicum, Episyrphus balteatus, Exorista larvarum, Glossina morsitans, Lonchoptera uniseta, Musca domestica, Minettia flaveola, Megaselia scalaris, Sarcophaga bullata, Sepsis*

Table 3.1. Species list and family level identification

In the case of different data sources for mitochondrial and nuclear gene sequences, the specific species are listed in parenthesis, with the mitochondrial data source appearing first and the nuclear data source appearing second.

Taxon	Family
Asilus crabroniformis	Asilidae
Anopheles gambiae	Culicidae
Bombylius major	Bombyliidae
Ceratitis capitata	Tephritidae
Cyrtodiopsis dalmanni	Diopsidae
Cochliomyia sp. (*C. hominivorax* / *C. macellari*)	Calliphoridae
Drosophila melanogaster	Drosophilidae
Delia radicum	Anthomyiidae
Episyrphus balteatus	Syrphidae
Exorista larvarum	Tachinidae
Empididae sp. (*E. barbatoides* / *O. scopifer*)	Empididae
Glossina morsitans	Glossinidae
Haematopota pluvialis	Tabanidae
Lonchoptera uniseta	Lonchopteridae
Musca domestica	Muscidae
Minettia flaveola	Lauxaniidae
Megaselia scalaris	Phoridae
Sarcophaga bullata	Sarcophagidae
Sepsis cynipsea	Sepsidae
Scatophaga stercoraria	Scatophagidae

cynipsea, and *Scatophaga stercoraria* were obtained via the above method. Additional mitochondrial gene sequences and nuclear gene data was obtained from the FLYTREE project group (Wiegmann et al. 2011).

Data matrix preparation

Single gene alignments using the translated amino acid sequence were performed with MEGA 4.0 (Kumar et al. 2008) to produce a codon alignment based on translated protein sequence. Variable sites and regions of poor alignment were removed using Gblocks (Talavera, Castresana 2007) in codon mode with default block parameters and a 50% missing sites threshold. All thirteen protein coding genes from the mitochondrial genome were concatenated to produce an alignment of 11,217 base pairs in length. After removing highly variable and poorly represented sites, the resulting mitochondrial alignment included 10,425 base pairs. Twelve protein encoding nuclear genes were selected for analysis, with a total combined length of 11,946 bases. The entire sequence of two of the genes (*pug* and *stx*) was removed due to a failure to identify any conserved blocks with Gblocks. This left ten genes totaling 7,770 base pairs in length in the nuclear gene alignment. In addition to the mitochondrial and nuclear gene alignments, a concatenated alignment was created. The concatenated sequence of mitochondrial and nuclear genes contained twenty three genes and 18,195 base pairs (Table 3.2).

Table 3.2. Sequence length statistics

Number of non-ambiguous sites recovered for each gene in each taxon. The number prior to the slash indicates the total number of sites. The number following the slash indicates the number of sites remaining after using Gblocks (Castresana 2000) to trim poorly represented and highly variable sites from the alignment. The Gblocked alignments were used for all analyses. A dash indicates that a sequence was not recovered or had zero sites remaining after Gblocks. Genes in bold are nuclear genes excluded from analyses due to the fact that they contained zero sites after variable sites were removed with Gblocks.

Mitochondrial genes

Taxon	*atp6*	*atp8*	*cox1*	*cox2*	*cox3*	*cytb*	*nd1*	*nd2*	*nd3*	*nd4*	*nd4l*	*nd5*	*nd6*
Asilus crabroniformis	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1023 / 879	348 / 345	1338 / 1302	288 / 273	1713 / 1704	519 / 123
Anopheles gambiae	672 / 666	153 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1020 / 879	348 / 345	1341 / 1302	294 / 273	1713 / 1704	519 / 123
Bombylius major	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1026 / 879	348 / 345	1338 / 1302	285 / 273	1713 / 1704	519 / 123
Ceratitis capitata	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1017 / 879	348 / 345	1338 / 1302	288 / 273	1713 / 1704	519 / 123
Cyrtodiopsis dalmanni	669 / 666	- / -	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	- / -	348 / 345	1338 / 1302	- / -	1706 / 1697	- / -
Cochliomyia sp. (*C. hominivorax* / *C. macellari*)	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1011 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	516 / 123
Drosophila melanogaster	666 / 663	153 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1020 / 879	348 / 345	1338 / 1302	288 / 273	1719 / 1704	519 / 123
Delia radicum	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1011 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	516 / 123
Episyrphus balteatus	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1019 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	519 / 123
Exorista larvarum	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1011 / 879	348 / 345	1350 / 1302	- / -	1710 / 1701	- / -
Empididae sp. (*E. barbatoides* / *O. scopifer*)	669 / 666	153 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1032 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	519 / 123
Glossina morsitans	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1002 / 879	348 / 345	1338 / 1302	288 / 273	1713 / 1704	513 / 123
Haematopota pluvialis	669 / 666	153 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1038 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	519 / 123
Lonchoptera uniseta	669 / 666	- / -	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	941 / 785	348 / 345	1347 / 1302	294 / 273	1713 / 1704	516 / 123
Musca domestica	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1005 / 876	348 / 345	1338 / 1302	294 / 273	1713 / 1704	516 / 123
Minettia flaveola	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1020 / 873	348 / 345	1338 / 1302	294 / 273	1713 / 1704	519 / 123
Megaselia scalaris	669 / 666	- / -	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1022 / 876	348 / 345	1335 / 1302	285 / 273	1713 / 1704	- / -
Sarcophaga bullata	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1010 / 879	348 / 345	1338 / 1302	294 / 273	1713 / 1704	516 / 123
Sepsis cynipsea	669 / 666	153 / 108	1530 / 1530	678 / 666	783 / 780	1131 / 1128	930 / 921	1041 / 876	348 / 345	1338 / 1302	265 / 244	1709 / 1700	- / -
Scatophaga stercoraria	669 / 666	156 / 108	1530 / 1530	681 / 666	783 / 780	1131 / 1128	930 / 921	1019 / 879	348 / 345	1350 / 1302	288 / 273	1713 / 1704	516 / 123

Nuclear genes

Taxon	*aats1*	*aats2*	*cad*	*g6pd*	*pepck*	*per*	*pgd*	*pug*	*sia*	*snf*	*stx*	*tpi*
Asilus crabroniformis	813 / 753	665 / 642	3926 / 3317	697 / 447	458 / 366	615 / 588	771 / 522	- / -	441 / 405	342 / 252	- / -	483 / 453
Anopheles gambiae	813 / 753	1722 / 642	3951 / 3309	743 / 447	458 / 366	678 / 582	771 / 522	**637** / -	441 / 405	345 / 252	**474** / -	465 / 453
Bombylius major	779 / 744	1715 / 642	3834 / 3309	- / -	- / -	667 / 583	- / -	**637** / -	345 / 345	342 / 252	- / -	468 / 453
Ceratitis capitata	789 / 753	1682 / 642	3914 / 3323	741 / 444	455 / 366	686 / 603	785 / 522	**637** / -	438 / 405	345 / 252	**573** / -	482 / 452
Cyrtodiopsis dalmanni	801 / 753	1722 / 642	3912 / 3326	741 / 444	425 / 366	684 / 603	705 / 492	**637** / -	354 / 354	339 / 252	- / -	483 / 453
Cochliomyia sp. (*C. hominivorax* / *C. macellari*)	783 / 738	1040 / 14	3933 / 3327	741 / 444	454 / 366	684 / 600	784 / 522	**622** / -	417 / 393	336 / 252	**573** / -	475 / 453
Drosophila melanogaster	813 / 753	1722 / 642	3939 / 3327	744 / 447	455 / 366	687 / 603	771 / 522	**637** / -	435 / 405	345 / 252	**573** / -	483 / 453
Delia radicum	810 / 752	572 / 566	2394 / 2323	669 / 444	449 / 366	677 / 598	- / -	**636** / -	441 / 405	338 / 252	**573** / -	- / -
Episyrphus balteatus	802 / 752	1268 / 629	3902 / 3326	735 / 444	458 / 366	627 / 603	766 / 521	**619** / -	441 / 405	330 / 248	**573** / -	483 / 453
Exorista larvarum	812 / 753	642 / 630	3916 / 3326	705 / 444	455 / 366	684 / 600	780 / 522	**637** / -	441 / 405	337 / 252	**560** / -	444 / 426
Empididae sp. (*E. barbatoides* / *O. scopifer*)	555 / 552	671 / 642	3929 / 3326	744 / 447	437 / 366	619 / 597	755 / 522	- / -	348 / 342	247 / 227	- / -	479 / 453
Glossina morsitans	633 / 609	1277 / 642	2897 / 2726	473 / 444	451 / 366	605 / 592	746 / 522	- / -	441 / 405	261 / 240	- / -	483 / 453
Haematopota pluvialis	583 / 574	1339 / 642	3483 / 3296	693 / 447	458 / 366	616 / 592	536 / 495	**601** / -	410 / 400	- / -	- / -	483 / 453
Lonchoptera uniseta	807 / 753	1722 / 642	3897 / 3327	731 / 447	- / -	689 / 602	774 / 522	**637** / -	438 / 405	267 / 246	**453** / -	483 / 453
Musca domestica	799 / 752	1370 / 642	3822 / 3309	692 / 444	440 / 366	684 / 600	753 / 522	**637** / -	420 / 405	342 / 252	**574** / -	462 / 444
Minettia flaveola	813 / 753	665 / 642	3915 / 3327	674 / 444	392 / 363	686 / 602	771 / 522	- / -	441 / 405	336 / 252	**489** / -	483 / 453
Megaselia scalaris	764 / 753	1719 / 642	3897 / 3321	722 / 447	455 / 366	690 / 603	783 / 522	- / -	441 / 405	336 / 252	**576** / -	438 / 436
Sarcophaga bullata	777 / 753	220 / 213	3870 / 3327	741 / 444	455 / 366	684 / 600	786 / 522	**637** / -	429 / 405	345 / 252	**573** / -	477 / 453
Sepsis cynipsea	800 / 753	1677 / 637	2230 / 2161	- / -	378 / 352	647 / 603	577 / 522	**637** / -	417 / 405	341 / 252	**573** / -	483 / 453
Scatophaga stercoraria	799 / 753	1277 / 642	3868 / 3325	700 / 444	454 / 366	684 / 600	- / -	**637** / -	354 / 354	337 / 245	**573** / -	479 / 453

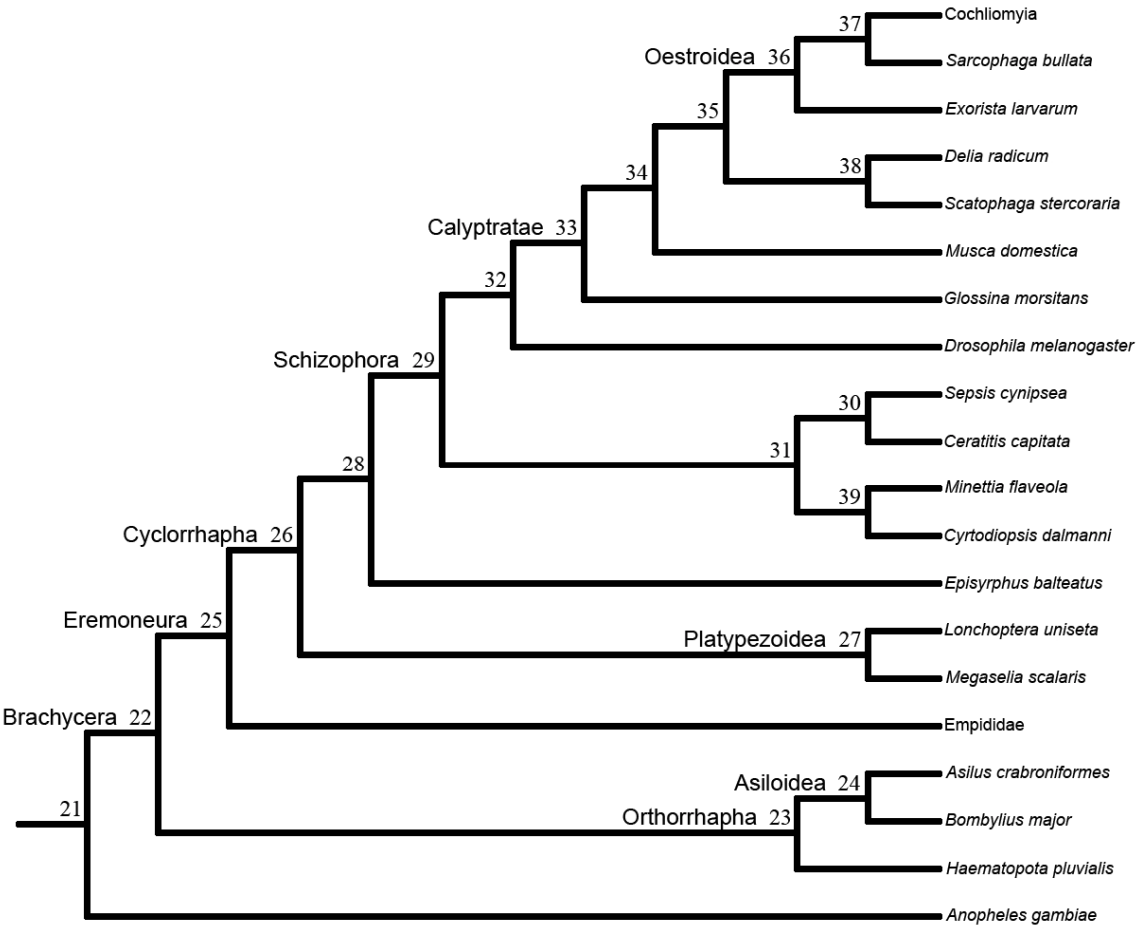
Two different partitioning schemes were applied to all data sets. For one set of analyses, data partitions were created for each gene, containing all included codon positions for a given gene within a single partition. Using this method, 13 mitochondrial and 10 nuclear gene partitions were created. A second set of data files was created that was partitioned based only upon codon position and data source (mitochondrial or nuclear), containing all data for a single codon position from all mitochondrial or nuclear genes within a single partition. This resulted in separate partitions three partitions for nuclear genes and three partitions for mitochondrial genes.

Divergence time estimation

Divergence time estimation was performed using the BEAST 1.6.1 (Drummond, Rambaut 2007). Tree topology was fixed to the topology arrived at by the FLYTREE project (Wiegmann et al. 2011) (Fig 3.2); however, we transposed the position of Sarcophagidae and Tachinidae as our analyses recovered a Sarcophagidae/Calliphoridae clade exclusively (not shown). Each data partition was assigned an independent substitution model, either HKY or GTR with both a four category gamma site heterogeneity model and an invariant sites parameter. A shared relaxed clock model (uncorrelated lognormal) was linked to all partitions, as was a shared Yule process speciation tree model. All data sets were run for 1 million generations at least five consecutive times to optimize model parameters prior to the final run of 20 million generations. For both the tuning runs and the final run, the trees were sampled every 200 generations. Tracer 1.5 (Rambaut, Drummond 2007) was used to analyze the BEAST log files.

Figure 3.2. Tree topology and clade numbering

Fixed tree topology used in all clade age calculations with each clade numbered. Selected clade names appear to the left of the corresponding node number. Tree topology adapted from Wiegmann *et al.* (2011)



Fossil calibration

Upper and lower age boundaries were selected based on the available fossil evidence to calibrate the tree. The root height was calibrated to 210-230my (*Aenne – Grauvogelia*) (Grimaldi, Engel 2005). Brachycera was calibrated 195-210my (*Oligophyrne*) (Grimaldi, Engel 2005). Cyclorrhapha was set to 125-135my (*Opetiala*) (Grimaldi, Engel 2005) and Schizophora was set to 64-74 my (*Phytomyzites*) (Winkler et al. 2010). Using only these calibration points, preliminary age estimates were much younger for many schizophran clades than could be justified by the fossil record (not shown). Thus an age range of 42-52my was assigned to the Anthomyiidae/Scatophagidae split (*Protanthomyia*) (Grimaldi, Engel 2005) to compress the schizophoran radiation to match available fossil data.

To model these age ranges within BEAST, it was necessary to assign a prior distribution to these nodes. For each node, we assigned a normal distribution with mean equivalent to the middle of the expected age range and a standard deviation was selected such that 80% of the distribution fell within the expected age range (Table 3.3). For each node, less informative wider distributions were also tested. These more permissive priors, however, allowed BEAST to infer unrealistic ages for the calibrated nodes, which led us to conclude that their performance was inferior to the more strictly enforced calibration point. As the shift in estimated ages towards ages not supported by the fossil record got progressively more severe as the strength of the prior was weakened from 90% of the distribution falling within the expected range down to only 40%, we selected the 80% category as a compromise to maintain strict calibration while still allowing flexibility for the data to influence the results of our calibration points.

Table 3.3. Fossil calibration distributions

Fossil calibration data showing fossil age and estimated range of fossil calibration. Median and standard deviation values were calculated such that 80% of the resulting normal distribution would lie between the estimated minimum and maximum age for the clade.

Clade	Fossil	Est. age range	Median	Std. dev.
Root	210	230-210	220	7.803045
Brachy	195	210-195	202.5	5.852284
Cyclo	125	125-135	130	3.901522
Schizo	64	64-74	69	3.901522
Antho	42	52-42	47	3.901522

Analysis of ESS

In order to compare the ability of a given data set and model to resolve a clade and select which is performing better, we looked at the effective sample size (ESS) of the clade age as derived from the BEAST trace files. ESS represents the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to. While ESS is not a direct estimator of confidence, it is an indication of how well the node is being sampled by the algorithm given the evolutionary models, clock model, tree topology, and data set. ESS's can differ from one program run to the next, although they are generally similar between successive analyses. Lower ESS's indicate poor sampling of the node due to high correlation between samples and relatively poorer performance than a higher sample size. Low ESS can be directly overcome by increasing the length of the analysis or by increasing the sampling frequency. As our focus was on the information content of the genes and the relative merits of altering the models or data set composition, we fixed the number of generations and sampling frequency. As suggested by the BEAST documentation, we chose 100 ESS as the lower cutoff for moderate confidence in a result, with any node falling below 100 ESS in a given analysis being considered to have too poor of a sampling to give a highly reliable estimate of clade age. Furthermore, we considered the threshold category composed of nodes for which the ESS fell between 100 and 200 to be clades for which inference is difficult and misestimations due to insufficient sampling are possible.

Results

Sequence comparison

As expected, the mitochondrial and nuclear encoded genes displayed notably different patterns of sequence evolution. In addition to the decreased number of conserved amino acid sites present in the nuclear gene data relative to the mitochondrial gene data, average base composition, the degree of species specific deviation from the average, and 3rd codon substitution patterns varied dramatically between data sets (Fig 3.3, Table 3.4).

Average base composition for the mitochondrial genes was 31.39% A, 12.39% C, 13.07% G, and 43.15% T. All taxa except for the hornet robberfly *Asilus crabroniformes* fell within ±2.32% of the average. In *Asilus*, a substitution bias of nearly 7% favoring C over T and nearly 4% favoring G over A compared to the average base composition was observed. With removal of 3rd codon positions, variation between base frequencies was less than ±1.84%, and in the case of *Asilus* the bias shrank to 3.88% and 1.69% respectively. Average base frequencies for the nuclear genes were 28.76% A, 20.37% C, 23.88% G, and 26.99% T with variations of up to 9.94% from the mean base frequency observed in some taxa. Removal of 3rd codon positions dramatically reduced the variations in base composition with a maximum variation of ±3.42% observed.

Overall, mitochondrial genome encoded genes had base frequencies strongly skewed in favor of AT but showed little species specific deviation from the average. Furthermore, the species specific variations in base frequency were concentrated in 3rd codon positions. Removal of 3rd codon positions lessened the AT bias; however, base compositions were still skewed. The taxon *A. crabroniformes* showed a notably weaker AT bias in its mitochondrial genome than any other included taxon, and this affected 1st and 2nd codon positions as well as 3rd. Nuclear encoded genes, on the other hand, had

Figure 3.3. Base composition of mitochondrial and nuclear genes

Shaded bars represent the average frequency over all species for that base. Error bars indicate standard deviation. All comparisons between mitochondrial and nuclear genes showed statistically significant differences in base frequencies (two tailed t-test, unequal variances).

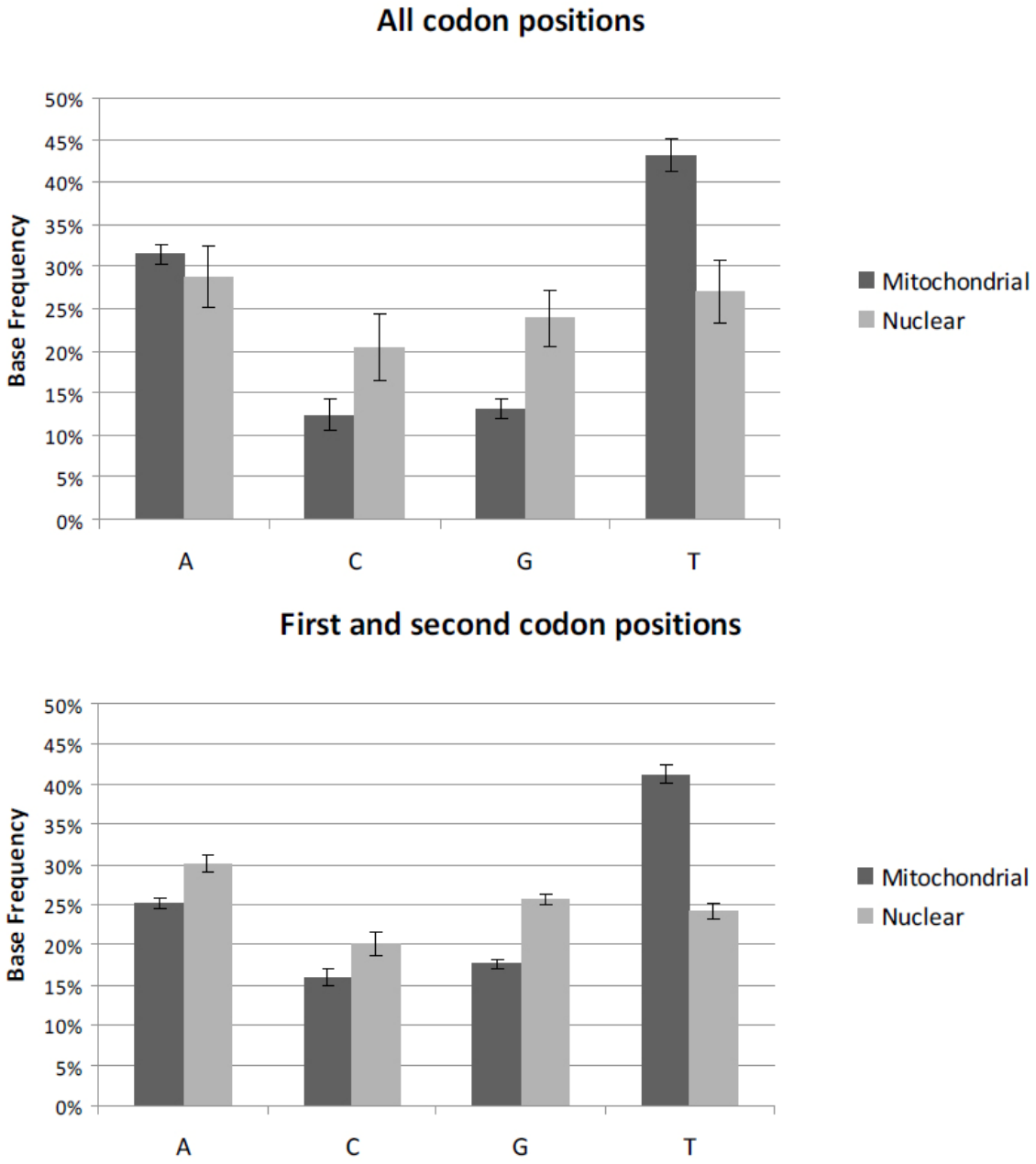


Table 3.4. Average base composition

Mitochondrial and nuclear gene base compositions calculated both with and without 3rd codon position data. Sites column represents the total number of nucleotide sites used to calculate the averages for that species. Base frequencies represent the amount of divergence relative to the average base composition calculated across all taxa.

Mitochondrial Gene Base Composition

	All positions					3rd codon excluded				
Taxon	Sites	A	C	G	T	Sites	A	C	G	T
Asilus crabroniformis	10425	-3.54%	6.63%	3.81%	-6.90%	6950	-1.44%	3.63%	1.69%	-3.88%
Anopheles gambiae	10425	0.30%	-0.46%	-0.14%	0.30%	6950	-0.32%	-0.23%	0.02%	0.52%
Bombylius major	10425	-1.51%	2.32%	1.33%	-2.14%	6950	-0.73%	1.67%	0.90%	-1.84%
Ceratitis capitata	10425	-0.45%	-0.24%	-0.19%	0.88%	6950	-0.45%	0.07%	0.13%	0.25%
Cyrtodiopsis dalmanni	9029	0.49%	-1.13%	-0.23%	0.87%	6019	0.78%	-1.01%	-0.24%	0.47%
Cochliomyia sp. (*C. hominivorax*)	10425	-0.03%	0.23%	0.15%	-0.36%	6950	-0.06%	-0.04%	-0.09%	0.19%
Drosophila melanogaster	10422	0.54%	-1.35%	-0.69%	1.50%	6948	-0.02%	-0.84%	-0.36%	1.22%
Delia radicum	10425	0.54%	-0.59%	-0.67%	0.73%	6950	0.01%	-0.30%	-0.45%	0.74%
Episyrphus balteatus	10422	1.88%	-1.92%	-1.69%	1.72%	6948	0.90%	-1.12%	-1.19%	1.41%
Exorista larvarum	8712	1.24%	-1.48%	-1.23%	1.47%	5808	0.74%	-0.65%	-0.57%	0.48%
Empididae sp. (*E. barbatoides*)	10425	-0.75%	1.51%	0.45%	-1.20%	6950	0.37%	0.33%	-0.32%	-0.39%
Glossina morsitans	10424	0.19%	-0.67%	-0.83%	1.32%	6949	0.44%	-0.44%	-0.75%	0.76%
Haematopota pluvialis	10236	0.07%	0.01%	0.23%	-0.31%	6826	-0.25%	0.23%	0.33%	-0.32%
Lonchoptera uniseta	9824	0.95%	-1.64%	-0.67%	1.36%	6550	0.76%	-1.07%	0.15%	0.16%
Musca domestica	10422	0.48%	-0.63%	-0.31%	0.46%	6948	-0.19%	-0.24%	-0.09%	0.52%
Minettia flaveola	10268	-0.83%	0.27%	0.10%	0.46%	6845	-0.83%	0.44%	0.24%	0.15%
Megaselia scalaris	9605	1.25%	-0.97%	-0.75%	0.47%	6404	1.06%	-0.82%	-0.58%	0.35%
Sarcophaga bullata	10425	-0.36%	0.89%	0.44%	-0.97%	6950	-0.32%	0.58%	0.10%	-0.36%
Sepsis cynipsea	9113	-0.84%	0.05%	1.23%	-0.43%	6075	-0.30%	-0.08%	0.84%	-0.46%
Scatophaga stercoraria	9572	0.37%	-0.83%	-0.32%	0.78%	6381	-0.16%	-0.11%	0.23%	0.04%
Average		**31.39%**	**12.39%**	**13.07%**	**43.15%**		**25.22%**	**15.95%**	**17.60%**	**41.22%**

Nuclear Gene Base Composition

	All positions					3rd codon excluded				
Taxon	Sites	A Avg	C Avg	G Avg	T Avg	Sites	A	C	G	T
Asilus crabroniformis	7752	0.88%	0.45%	-0.05%	-1.28%	5168	0.90%	-0.17%	0.00%	-0.73%
Anopheles gambiae	7731	-8.13%	7.80%	9.52%	-9.19%	5154	-2.24%	3.42%	1.43%	-2.61%
Bombylius major	6328	6.64%	-5.85%	-4.80%	4.01%	4219	1.71%	-1.08%	-0.97%	0.34%
Ceratitis capitata	7763	-0.90%	1.12%	0.70%	-0.92%	5175	-0.81%	0.81%	0.43%	-0.42%
Cyrtodiopsis dalmanni	7686	2.69%	-2.45%	-2.24%	2.01%	5124	0.18%	-0.33%	-0.27%	0.42%
Cochliomyia sp. (*C. macellari*)	7110	0.22%	-1.59%	-1.34%	2.70%	4740	0.21%	-0.94%	-0.28%	1.01%
Drosophila melanogaster	7770	-6.59%	7.84%	5.48%	-6.73%	5180	-1.57%	2.86%	1.08%	-2.37%
Delia radicum	5710	1.97%	-2.05%	-1.61%	1.69%	3807	1.00%	-1.34%	-0.10%	0.44%
Episyrphus balteatus	7763	0.37%	-1.10%	-0.56%	1.28%	5175	-0.01%	-0.13%	0.22%	-0.08%
Exorista larvarum	7725	0.75%	-1.23%	-1.16%	1.65%	5150	0.46%	-0.78%	-0.34%	0.66%
Empididae sp. (*O. scopifer*)	7475	1.69%	-2.10%	-0.81%	1.22%	4984	-0.29%	-0.01%	-0.15%	0.45%
Glossina morsitans	6999	0.82%	-1.29%	-1.13%	1.61%	4666	0.55%	-0.44%	-0.57%	0.46%
Haematopota pluvialis	7265	1.41%	-1.70%	-0.57%	0.86%	4843	-0.04%	-0.51%	-0.30%	0.85%
Lonchoptera uniseta	7404	3.33%	-3.89%	-2.56%	3.12%	4936	0.69%	-1.22%	-0.50%	1.03%
Musca domestica	7737	-1.08%	0.31%	0.48%	0.29%	5158	-0.14%	-0.77%	0.32%	0.59%
Minettia flaveola	7764	1.02%	-0.87%	-1.24%	1.09%	5176	0.03%	0.25%	-0.31%	0.03%
Megaselia scalaris	7747	1.63%	-0.64%	-2.20%	1.21%	5165	0.56%	-0.79%	-0.31%	0.53%
Sarcophaga bullata	7335	0.58%	-1.33%	-1.29%	2.04%	4890	0.05%	-0.88%	-0.22%	1.05%
Sepsis cynipsea	6142	-7.84%	9.94%	6.33%	-8.43%	4094	-1.90%	2.95%	1.04%	-2.09%
Scatophaga stercoraria	7191	0.53%	-1.36%	-0.95%	1.78%	4794	0.68%	-0.90%	-0.20%	0.42%
Average		**28.76%**	**20.37%**	**23.88%**	**26.99%**		**30.09%**	**20.07%**	**25.66%**	**24.18%**

average base frequencies which were more nearly equal, but which showed a high degree of variation among taxa. Removal of 3rd codon positions had little effect on average base frequencies, although it did reduce species specific deviation from the average. The taxa *Anopheles gambiae*, *Drosophila melanogaster*, and *Sepsis cynipsea* showed the largest deviations from the average nuclear gene base frequencies and retained much of their variation even when 3rd codon positions were excluded from the data set.

Mitochondrial and nuclear divergence time estimates converge

In order to investigate the performance of mitochondrial genome encoded genes versus nuclear genome encoded genes for divergence time estimation, identical analyses were carried out on both data sets. Performance was assessed by comparing mean values and confidence intervals of divergence time estimates and by analyzing ESS support per node between mitochondrial and nuclear results. Data was partitioned by gene with each data partition containing all first and second codon positions for that gene and an HKY model assigned to each partition. For the majority of nodes in the tree, analysis of mitochondrial and nuclear genes produced age estimates within five million years of each other (Table 3.5). There were four notable exceptions to this. The mitochondrial gene data produced an age 52 million years younger than the nuclear gene data for the age estimates of the Asiloidea clade (node 24). The Platypezoidea clade (node 27),estimates differed by 24 my between the data sets, with mitochondrial gene data producing the younger estimate. For node 28, which unites the syrphids to the Schizophora, the estimates produced from the mitochondrial data set were 13 my younger than estimates from the nuclear data set. Node 39, which represents the split between Minettia and Cyrtodiopsis in our tree, was six my older in the mitochondrial estimate.

Table 3.5. Divergence times using gene based partitions

Divergence time estimates derived from a data set where separate partitions were assigned to each gene. Node labels correspond to the node labeling in Fig 3.2. Each estimate is displayed as median age in millions of years followed by the bounds of its 95% confidence interval. Node ages in red had ESS's below 100. Node ages in yellow had ESS's below 200. Data sets labeled with an asterisk had less than 100 ESS for the overall posterior probability.

		HKY Model			GTR Model		
node	clade name	mitochondrial, 3rd codon excl.	nuclear, 3rd codon excl.	concatenated, 3rd codon excl.	mitochondrial, 3rd codon excl.*	nuclear, 3rd codon excl.*	concatenated, 3rd codon excl.*
21		219.1052 (233.8523, 204.5721)	219.18 (233.6459, 204.6366)	219.3035 (233.9009, 205.0084)	219.2758 (233.9552, 204.8379)	219.3379 (233.8477, 204.6946)	219.3364 (234.3566, 204.7533)
22	Brachycera	199.1993 (210.4671, 188.0299)	198.8579 (209.9255, 187.469)	198.8423 (210.0182, 187.3646)	199.1285 (210.2911, 187.8155)	198.7012 (209.9335, 187.5131)	198.1283 (209.5889, 186.7655)
23	Orthorrhapha	190.8295 (209.4003, 155.3812)	190.863 (207.5991, 166.3969)	190.5175 (208.9165, 163.0297)	192.8439 (210.1829, 165.9143)	192.7753 (208.3026, 170.715)	194.0715 (208.6562, 177.619)
24	Asiloidea	113.0346 (175.9937, 43.3605)	165.3646 (198.0025, 99.7499)	27.0771 (41.6391, 13.4907)	124.5353 (182.4462, 60.2291)	164.6326 (198.8788, 92.2362)	131.0274 (178.2351, 21.6328)
25	Eremoneura	169.5605 (199.9223, 134.8033)	170.3342 (196.108, 140.8415)	172.457 (198.5163, 139.1587)	169.7578 (199.5481, 133.6036)	169.4097 (195.1058, 139.5653)	173.4 (198.171, 140.2857)
26	Cyclorrhapha	130.2907 (138.1338, 123.0014)	130.1035 (137.4082, 122.5959)	130.1979 (137.5018, 122.5293)	130.3263 (137.9843, 122.8677)	130.084 (137.4354, 122.4936)	130.3521 (137.817, 122.8805)
27	Platypezoidea	96.0004 (127.66, 43.5625)	120.0976 (134.1859, 96.3252)	103.9844 (129.2839, 63.941)	97.2204 (127.0459, 51.4786)	119.4078 (134.0613, 93.8975)	110.5353 (130.2282, 68.4085)
28		89.4978 (117.166, 72.7863)	101.9436 (121.521, 83.4887)	98.4766 (120.9726, 78.2956)	88.4707 (115.9048, 73.0008)	101.623 (121.0468, 83.0132)	97.3088 (119.7148, 78.679)
29	Schizophora	74.9577 (81.9418, 67.73)	76.0737 (83.19, 69.3227)	75.4235 (82.5839, 68.7252)	74.793 (81.8984, 67.6914)	76.0731 (83.1505, 68.9889)	76.0037 (83.5228, 68.3971)
30		51.4066 (70.5775, 26.6702)	54.6593 (70.6998, 33.1173)	52.9555 (70.6589, 28.827)	53.7217 (71.0477, 28.1183)	56.057 (71.3359, 36.6957)	55.9929 (72.1414, 35.4876)
31		66.2968 (78.7986, 48.3114)	68.5764 (79.2727, 55.2011)	67.1849 (78.3722, 51.4983)	66.8005 (78.2061, 49.4107)	68.5866 (79.305, 55.0211)	68.7437 (79.5956, 54.8796)
32		69.2494 (77.219, 60.6222)	71.6705 (79.4059, 63.9098)	70.2002 (77.853, 62.6992)	69.1422 (77.1198, 60.9065)	71.6695 (79.2456, 63.5746)	70.9093 (79.2221, 62.0461)
33	Calyptratae	59.1652 (68.1132, 50.2814)	60.3275 (68.2082, 52.2284)	59.6815 (67.4119, 51.9942)	59.0963 (67.5374, 50.1621)	59.9386 (67.4861, 52.1859)	59.8289 (67.9486, 51.1421)
34		51.0909 (60.0772, 42.7882)	53.7429 (61.4876, 45.9159)	52.6795 (60.1672, 45.2715)	50.9991 (59.723, 42.2002)	53.5588 (61.0802, 45.9399)	52.1705 (60.162, 44.1502)
35		45.9233 (53.8728, 38.3531)	46.5789 (53.9428, 39.3836)	46.6709 (53.9956, 39.7025)	46.0358 (54.0434, 37.7616)	46.5808 (53.9369, 39.6493)	46.0674 (53.649, 39.1728)
36	Oestroidea	38.7016 (50.2334, 23.5625)	34.5626 (45.8523, 21.2792)	37.3226 (47.5263, 21.686)	38.8557 (49.9153, 25.2998)	33.7039 (44.7225, 20.9004)	36.3493 (46.957, 23.8712)
37		26.7233 (40.6375, 9.8774)	22.4234 (35.4779, 10.1421)	24.3485 (37.879, 9.4298)	26.6336 (41.3042, 10.9506)	21.9302 (34.1759, 9.9602)	24.4036 (36.3256, 10.3976)
38		41.0889 (47.9545, 34.3966)	40.039 (46.3806, 33.3996)	40.9166 (47.5837, 34.5029)	41.1833 (47.8666, 33.7925)	40.1742 (46.6448, 33.5012)	40.5139 (47.4899, 34.0279)
39		60.2482 (75.8014, 36.1117)	53.5022 (70.3937, 33.787)	53.7671 (71.1085, 32.3237)	61.3976 (75.6964, 41.4418)	54.2253 (70.1873, 33.3851)	58.8231 (74.6849, 37.6738)

		HKY Model			GTR Model		
node	clade name	mitochondrial, all sites	nuclear, all sites	concatenated, all sites	mitochondrial, all sites*	nuclear, all sites*	concatenated, all sites*
21		219.0192 (233.9122, 204.7999)	219.3685 (234.3316, 205.0657)	219.1861 (234.0542, 204.6986)	219.0036 (233.9698, 204.7605)	219.425 (234.1537, 204.7032)	219.2714 (233.9287, 204.632)
22	Brachycera	198.769 (210.0282, 187.3991)	197.9315 (209.0564, 186.2679)	197.3421 (208.5096, 185.706)	198.658 (210.0513, 187.2506)	197.9749 (209.2858, 186.2645)	197.8972 (209.7413, 186.0811)
23	Orthorrhapha	189.7164 (207.1557, 163.5026)	188.6815 (205.9615, 163.1411)	189.6958 (204.9317, 170.8278)	192.8508 (208.8865, 167.881)	191.1669 (207.5649, 168.0368)	193.2376 (207.8069, 177.6509)
24	Asiloidea	117.5786 (169.0396, 59.4124)	167.9935 (196.3116, 121.1873)	137.2949 (181.8515, 83.9375)	118.333 (176.4607, 49.7839)	174.6427 (200.0674, 130.1185)	138.8298 (190.3091, 68.784)
25	Eremoneura	157.1793 (194.1841, 130.9046)	174.2004 (197.6965, 143.2037)	175.309 (199.3765, 144.5882)	157.1791 (193.5557, 130.4911)	175.0798 (198.6126, 142.7234)	176.1126 (199.5837, 144.0404)
26	Cyclorrhapha	129.95 (137.5803, 122.421)	130.2365 (137.6988, 122.7506)	130.9517 (138.2798, 123.3271)	130.3329 (137.9009, 122.709)	130.5153 (138.052, 123.1118)	130.4678 (137.9525, 122.9598)
27	Platypezoidea	27.1039 (114.8896, 10.6332)	119.7931 (134.5317, 93.1436)	111.0368 (130.8898, 75.768)	99.4444 (129.1485, 54.7305)	121.1059 (134.3837, 98.3918)	105.6883 (128.8736, 55.9533)
28		88.4518 (114.2111, 72.718)	102.4736 (121.547, 84.1663)	99.0218 (122.2403, 78.6248)	86.9176 (114.4873, 71.0785)	101.9686 (122.7306, 81.4993)	98.6301 (121.2508, 79.124)
29	Schizophora	75.2493 (82.6049, 68.2606)	76.3242 (83.3099, 69.2168)	75.549 (82.8454, 68.4639)	74.6978 (82.0225, 67.291)	75.5386 (82.7624, 68.5778)	75.7066 (83.1567, 68.4762)
30		53.2416 (69.6461, 28.8981)	48.6768 (65.0728, 31.042)	50.7978 (67.3181, 30.608)	55.7175 (71.7979, 34.1154)	47.84 (63.6395, 29.6971)	52.9842 (68.2233, 23.9915)
31		65.8093 (77.9028, 47.9984)	68.8645 (78.7535, 56.879)	67.8705 (78.2135, 54.8804)	65.5196 (77.3472, 49.2401)	67.4192 (78.14, 53.6043)	68.3313 (79.6883, 54.1906)
32		70.1735 (77.8779, 61.8626)	71.5991 (79.125, 63.5664)	70.7874 (78.451, 62.7881)	69.7939 (77.6435, 61.1964)	70.6665 (78.3564, 62.5234)	70.8049 (79.1203, 62.2778)
33	Calyptratae	61.3549 (69.678, 52.3475)	61.9827 (69.8162, 54.1823)	61.6717 (71.0982, 52.8671)	61.2536 (69.714, 52.935)	61.3716 (69.4827, 53.3298)	61.5319 (70.1262, 52.2108)
34		54.1865 (62.7974, 45.5461)	55.3074 (63.2251, 47.6694)	54.609 (64.0659, 46.5512)	54.0209 (62.4963, 45.7503)	55.2523 (63.3937, 47.3804)	54.55 (63.408, 45.4091)
35		48.9199 (57.3017, 40.9609)	48.0765 (55.8072, 40.9231)	48.3805 (57.2139, 40.6423)	49.01 (57.036, 41.2448)	48.1881 (55.809, 41.2156)	47.9621 (56.7964, 40.0392)
36	Oestroidea	45.0805 (54.6488, 34.3525)	37.7209 (48.4021, 25.3645)	40.7523 (51.8686, 29.3021)	44.9579 (54.711, 33.9012)	37.7919 (48.3733, 25.5317)	40.2466 (51.6067, 28.4413)
37		31.9129 (43.7184, 16.2262)	25.7118 (37.7852, 13.1853)	29.3608 (40.7282, 16.2643)	31.7164 (45.261, 14.9374)	26.1121 (38.9646, 13.9073)	29.7691 (43.1975, 13.9925)
38		42.714 (49.4809, 35.8045)	41.0874 (47.7633, 34.1595)	41.6209 (48.6357, 34.7615)	42.9128 (49.3538, 35.9456)	41.4158 (47.7469, 34.9311)	41.4136 (48.0758, 34.6536)
39		58.2461 (73.6064, 32.8054)	52.9678 (67.6704, 33.7131)	53.9586 (68.9352, 34.4669)	58.6442 (73.8251, 39.1322)	52.4685 (68.4135, 29.5727)	57.3039 (73.8756, 34.4272)

All node age estimates except node 24 in the nuclear gene data set had ESS's in excess of 200. The mitochondrial gene data set, however, had six estimates which scored lower than 200 ESS (nodes 23, 24, 27, 30, 37 and 39) and one node that scored lower than 100 ESS (node 24). This indicated that under the model conditions and partitioning scheme used, the mitochondrial gene data was less effective at inferring divergence time information than the nuclear gene data set. Still, for most nodes the two sets of age estimates were remarkably close despite their very different evolutionary patterns and variations in ESS.

Concatenation of mitochondrial and nuclear gene data has a cost in computational complexity, but little benefit to accuracy

In the cases where we observed disagreement between estimates from mitochondrial and nuclear data sources, one data source may have contained a stronger signal for that node than the other. In order to test the relative signal strength in each data source, the data sets were concatenated. Analysis of the concatenated mitochondrial and nuclear gene data sets produced results very comparable to either mitochondrial or nuclear gene data alone (Table 3.5). For the nodes which showed disagreement between mitochondrial and nuclear gene derived estimates, the clade age estimates of the concatenated data set lay between the two estimates.

Overall, concatenation led to a decrease in ESS compared to the single data source partitions. Nine nodes fell below an ESS score of 200 and four of those were below 100. For most nodes (excluding the Asiloidea, node 24) with lower ESS relative to estimates derived from the nuclear encoded or mitochondrial genes alone, the decrease did not appear to have a noticeable adverse impact on divergence time estimates. It was,

however, indicative of an increase in computational complexity and an overall decrease in performance.

Inclusion of 3rd codon position data decreases consistency

Third codon positions are typically discarded when analyzing deeper level phylogenies due to high levels of homoplasy at these rapidly evolving sites. In our divergence time analyses, the tree topology was fixed, but homoplasy induced parameter misestimation was still likely to be an issue. To test whether increased data set size with the cost of increased homoplasy would have a negative impact on divergence time estimation, and if it did, whether it would be restricted to specific time depths, we ran a parallel set of analyses with 3rd codon data included to compare to 1st and 2nd codon position only results.

For the majority of nodes, inclusion of 3rd codon position data had little effect on the inferred age of the node (Table 3.5), nevertheless the exceptions indicated a probable negative effect on accuracy. When 3rd codon positions were included, the age estimate for node 27 derived from mitochondrial gene data fell by 69 my, resulting in a 107 my younger age than the estimate derived from nuclear gene data with either 3rd codon positions included or excluded. The Eremoneura clade age estimate (node 25) using mitochondrial gene data was 13 million years younger with the 3rd codon included, which caused it to fall out of agreement with the nuclear gene derived estimate. For node 30, inclusion of the 3rd codon position in the nuclear gene data set caused a six million year decrease in inferred age, reducing its level of agreement with mitochondrial estimate.

Further indicative of a negative effect, inclusion of the 3rd codon position reduced ESS's of both mitochondrial and nuclear data sets. For both 3rd codon included and 3rd

codon excluded mitochondrial data sets, six node ages were below 200 ESS. Four of those in the 3rd codon included data set were also below 100 ESS while only one was below that threshold in the 3rd codon excluded set. Within the nuclear gene derived estimates, 3rd codon inclusion caused the one estimate with lower than 200 ESS to fall below 100 (node 24), and node 37 to fall below 200 ESS. The effects on the concatenated data set were even more severe. Only eight of the nineteen nodes exceeding the 200 ESS required for adequate sampling and six nodes were below 100 ESS. As the number of parameters to estimate did not change with the inclusion of 3rd codon data, the most probable explanations for the loss of robustness was an increase in difficulty in fitting the model to the more complex and variable data set as well as the increased size of the data matrix.

A more complex model does not improve consistency

Our previous analyses using an HKY evolutionary model for all data partitions showed several nodes where estimates derived from either nuclear or mitochondrial encoded genes diverged. As it was possible that the simpler HKY model did not properly simulate the complexity of evolutionary patterns in one or both data sets and led to these discrepancies, a more parameter rich GTR model was tested on each data set.

In all but two cases, use of the more complex GTR model produced the same divergence date estimates as the simpler HKY model (Table 3.5). The only nodes and data sets for which use of the GTR model produced a substantially different result than the HKY model was node 24 in the concatenated 3rd codon position excluded data set and node 27 in the mitochondrial 3rd codon position included data set. In both cases, use of the GTR model produced a more reasonable estimate than the HKY model (131 my

rather than 27 my for node 24 and 99 my instead of 27 my for node 27), but low ESS values characterized these nodes under both HKY and GTR models.

Overall, using the GTR model had only a minor impact on ESS for most nodes. For 3rd codon excluded data sets, use of the more complex model slightly improved ESS values for both the mitochondrial and concatenated data sets, but had little impact on the nuclear gene derived estimates. The 3rd codon included data sets showed a different trend, with ESS values improving for mitochondrial gene data sets, but falling for concatenated and nuclear gene data sets.

Despite some minor improvements to node specific ESS values in some data sets, the overall ESS of the tree posterior fell dramatically. In all analyses performed with the GTR model, the overall ESS was below 100, and in most cases below 30. As predicted by earlier studies (Rannala 2002), use of the more parameter rich GTR model had a cost in computational complexity that would require analysis for a much longer period of time in order to obtain sample sizes similar those obtained using the HKY model.

Codon based partitioning produces similar results to gene based partitioning

Partitioning the data set by genes and assigning each gene an independent model is the obvious choice if one assumes that the difference in substitution patterns between genes is greater than the difference in patterns between 1st and 2nd codon positions within the same gene. Partitioning by gene, however, creates a greater number of smaller partitions in the data set that causes an increase in the number of parameters to estimate and a decrease in the amount of data available for the estimation of those parameters. In combination, those two factors can cause greater uncertainty in the results. In order to test a less parameter rich partitioning schema, we created data sets partitioned based only

upon codon position and data source (nuclear or mitochondrial). Each partition contained data from all related genes, but not from unrelated codon positions.

In 3rd codon excluded data sets, partitioning the data based on its codon position rather than by gene produced nearly identical results for every node (Table 3.6). The only exceptions to this were the inherently problematic Asiloidea node (node 24) where effective sampling in mitochondrial and concatenated data sets was typically so low that little confidence can be placed in the precision of any estimate, and the Platypezoidea clade (node 27) estimate produced from the concatenated data set under the HKY model. In this latter case, use of codon position based partitioning increased the age estimate by seven my and brought it into closer agreement with the estimates produced under the GTR model in both gene and codon position based partitioning analyses.

When 3rd codon positions were included, the differences between partitioning strategies became more obvious. While the age estimate derived from codon partitioned data produced inferior results for the Asiloidea clade (node 24) when used with mitochondrial sequence data, HKY model results for nodes 25 and 27 showed an improvement when analyzed with codon position partitioning. The median age estimate for node 25, for instance, increased from 157 my with gene based partitioning to 168 my with codon based partitioning. This was the highest degree of agreement with nuclear and concatenated data set results that we saw for this node among all other set of conditions analyzed. Similarly impressive, node 27 improved from an aberrantly low 27 my estimate with gene based partitioning to a more consistent 100 my estimate. Under the more complex GTR model, we saw no improvement in age estimation ability with the mitochondrial data when codon position based partitioning was used. There were,

Table 3.6. Divergence times using codon based partitions

Divergence time estimates derived from a data set where genes from the same source (mitochondrial or nuclear genome) were pooled and separate partitions were assigned to each codon position. Node labels correspond to the node labeling in Fig 3.2. Each estimate is displayed as median age in millions of years followed by the bounds of its 95% confidence interval. Node ages in red had ESS's below 100. Node ages in yellow had ESS's below 200. Data sets labeled with an asterisk had less than 100 ESS for the overall posterior probability.

		HKY Model			GTR Model		
node	clade name	mitochondrial, 3rd codon excl.	nuclear, 3rd codon excl.	concatenated, 3rd codon excl.	mitochondrial, 3rd codon excl.	nuclear, 3rd codon excl.	concatenated, 3rd codon excl.
21		219.1008 (233.6068, 204.8344)	219.3015 (234.0266, 204.9522)	219.4415 (234.4249, 204.7251)	219.2503 (234.2206, 205.1571)	219.1856 (234.1801, 205.0141)	218.8822 (233.5424, 204.133)
22	Brachycera	199.4804 (210.9207, 188.317)	198.6779 (209.9248, 187.3296)	198.3351 (209.8043, 186.7458)	199.3617 (210.5611, 188.0518)	198.7818 (209.8567, 187.3153)	198.8075 (209.9763, 187.0557)
23	Orthorrhapha	192.2471 (210.5957, 161.5252)	190.7326 (207.8335, 165.6636)	192.0583 (207.7899, 172.4607)	193.453 (211.5549, 165.6467)	192.4266 (208.5597, 167.9771)	194.9804 (208.9451, 179.2029)
24	Asiloidea	120.6767 (174.9846, 60.7425)	166.8914 (197.4567, 114.2029)	133.4577 (182.2747, 65.7595)	114.3165 (182.8699, 48.7874)	166.5245 (200.3155, 91.0996)	145.2718 (190.5654, 78.8438)
25	Eremoneura	164.0483 (197.5913, 132.4373)	171.8056 (198.0597, 142.6539)	172.086 (197.4186, 140.6909)	162.2786 (196.9169, 131.9562)	170.9158 (195.9884, 141.7206)	172.2744 (197.8722, 140.9348)
26	Cyclorrhapha	130.2572 (137.8266, 122.6767)	130.0912 (137.6717, 122.6296)	130.3361 (138.0223, 122.8716)	130.2144 (137.8174, 122.6929)	130.2299 (137.7973, 122.8642)	130.2304 (137.9784, 122.8371)
27	Platypezoidea	96.3152 (127.7113, 44.4729)	121.4465 (134.6977, 100.3414)	110.9996 (131.0299, 77.8545)	99.2915 (128.1194, 51.2227)	119.8358 (134.4397, 96.8558)	112.0898 (131.8457, 78.2358)
28		88.4072 (115.2414, 72.5825)	102.0153 (121.985, 82.8846)	97.1463 (118.9925, 79.0743)	87.8091 (114.4383, 72.0597)	101.8592 (120.9317, 82.6137)	97.4569 (119.0648, 78.8735)
29	Schizophora	74.7233 (81.836, 67.7746)	76.1403 (83.3676, 68.9057)	75.8088 (82.805, 68.4264)	74.8347 (81.9785, 67.6677)	75.9723 (83.185, 68.7171)	75.775 (82.9916, 68.7576)
30		52.5081 (70.4787, 28.9691)	54.8439 (69.7994, 33.8177)	54.7583 (71.3532, 31.9783)	52.9141 (71.186, 24.7118)	54.8008 (70.9315, 34.8557)	55.4772 (71.3771, 35.105)
31		67.3189 (78.7355, 49.6021)	68.7878 (79.3127, 55.2011)	68.7088 (79.5393, 53.858)	66.8833 (78.4396, 50.4532)	68.6192 (79.3286, 55.0737)	68.6078 (79.3033, 55.0614)
32		68.8604 (76.7091, 60.375)	71.5579 (79.5238, 63.753)	70.4129 (78.2047, 62.3216)	68.9281 (76.7971, 60.4066)	71.229 (79.2761, 63.1683)	70.3755 (78.4023, 62.2913)
33	Calyptratae	59.0008 (67.5136, 50.622)	59.9514 (67.9421, 52.2361)	59.4561 (67.9283, 51.4015)	58.9824 (67.9243, 50.6244)	59.6334 (67.217, 51.8454)	59.4716 (67.6716, 50.8591)
34		51.3259 (59.4787, 43.2253)	53.3085 (61.2078, 45.8224)	52.0373 (59.9169, 44.4847)	50.9962 (60.2196, 42.3045)	53.119 (60.9277, 45.6942)	52.1056 (60.4383, 44.0495)
35		45.9841 (53.7535, 38.1569)	46.2828 (53.882, 39.4157)	46.0571 (53.8275, 38.8987)	45.8308 (54.4354, 37.7417)	46.1096 (53.5518, 39.1128)	46.0943 (53.9173, 38.4018)
36	Oestroidea	38.0861 (49.2154, 23.3564)	34.4133 (45.9466, 21.9522)	36.808 (47.3117, 23.33)	38.405 (49.7923, 23.7566)	33.877 (45.0898, 19.5224)	37.2231 (48.1269, 25.2883)
37		28.0748 (42.2145, 12.247)	22.6722 (35.1028, 10.1221)	25.4738 (39.0484, 11.0843)	26.4879 (40.0336, 9.6072)	21.4734 (34.6639, 9.1151)	26.2034 (39.1765, 12.5225)
38		41.0668 (47.8805, 34.1619)	39.9474 (46.5071, 32.9888)	40.5704 (47.0616, 33.8927)	41.0131 (47.9633, 33.5806)	39.8324 (46.7022, 33.0882)	40.5241 (47.2467, 33.5971)
39		59.488 (75.1976, 37.8174)	53.3659 (69.3029, 30.0574)	55.8307 (73.137, 33.6692)	61.1786 (75.3991, 41.2189)	53.4344 (69.8246, 32.9604)	56.9801 (72.7457, 34.8854)

		HKY Model			GTR Model		
node	clade name	mitochondrial, all sites	nuclear, all sites	concatenated, all sites	mitochondrial, all sites*	nuclear, all sites*	concatenated, all sites*
21		218.8543 (233.6983, 204.4816)	219.3299 (234.3651, 204.9996)	218.9276 (234.362, 204.7179)	218.9452 (233.579, 204.5221)	219.4735 (234.2679, 204.9891)	219.4316 (234.481, 204.9483)
22	Brachycera	198.3529 (209.7048, 186.8362)	197.8336 (209.0021, 186.2436)	197.8243 (209.0357, 186.1624)	198.7169 (209.8035, 187.1717)	197.9204 (209.4695, 186.4614)	197.2219 (208.7453, 185.5787)
23	Orthorrhapha	190.7457 (208.6278, 159.0115)	190.558 (206.6855, 169.1706)	190.5643 (206.7109, 171.6241)	195.1835 (209.7068, 178.2808)	192.6576 (208.3932, 173.8816)	194.4921 (207.323, 181.2802)
24	Asiloidea	28.4929 (48.0671, 14.9892)	170.1573 (197.6134, 126.1095)	134.3912 (181.3445, 72.3164)	91.8644 (170.0624, 17.3546)	172.0342 (198.3457, 125.3435)	144.8245 (186.5315, 85.5722)
25	Eremoneura	168.3698 (199.4799, 134.08)	172.7774 (196.3494, 142.4702)	177.083 (199.8993, 146.4827)	160.9522 (195.6017, 131.5135)	169.8269 (196.6547, 138.9367)	168.9501 (193.5992, 140.4159)
26	Cyclorrhapha	130.4982 (138.0702, 122.9869)	130.4228 (138.0604, 122.9926)	130.5313 (137.965, 122.9675)	130.2603 (137.5719, 122.4572)	130.4151 (137.9845, 122.9343)	130.8461 (138.3649, 123.4936)
27	Platypezoidea	99.5613 (128.7003, 55.4453)	120.5812 (134.8256, 93.1194)	110.0377 (131.1224, 69.8805)	99.6605 (128.4152, 57.9364)	121.7234 (135.0479, 100.5409)	114.3748 (132.1089, 83.9346)
28		87.156 (115.8514, 72.0459)	104.2599 (122.7648, 85.954)	100.2369 (121.8922, 80.9701)	85.7463 (114.1815, 71.4183)	103.1915 (123.9386, 83.0564)	99.5638 (120.278, 81.0156)
29	Schizophora	75.0132 (82.1853, 68.0047)	76.2729 (83.1651, 69.3976)	75.9561 (83.1702, 69.182)	74.9438 (82.1043, 67.9432)	75.8115 (82.6139, 68.4463)	75.9546 (82.8556, 68.9725)
30		53.9689 (72.1315, 30.9145)	50.1342 (66.0368, 34.0143)	52.1395 (67.3119, 23.9253)	55.2617 (72.2038, 34.7039)	47.4546 (64.4104, 27.9406)	30.297 (65.1395, 17.7944)
31		66.187 (78.1327, 48.0163)	68.5259 (78.3562, 55.8008)	68.2458 (78.4367, 55.6868)	66.3786 (77.9457, 50.8933)	67.3719 (77.7369, 53.8124)	65.3882 (77.4646, 45.141)
32		69.6739 (77.4902, 61.1814)	71.378 (79.0179, 63.7004)	70.9112 (78.9152, 63.0443)	69.6797 (77.9784, 60.738)	70.423 (78.2014, 62.2479)	70.7107 (78.3039, 62.8194)
33	Calyptratae	60.9748 (69.5715, 52.1147)	61.5474 (69.3401, 53.7466)	62.1051 (70.1804, 54.0362)	61.4626 (70.3379, 52.859)	61.1682 (69.321, 52.9319)	61.0306 (69.384, 53.3404)
34		53.9221 (62.4779, 45.5521)	55.3193 (63.1014, 47.432)	55.3485 (63.6928, 47.2762)	54.272 (63.1799, 45.7093)	55.1099 (63.2798, 47.3471)	54.196 (62.2912, 46.9805)
35		48.5502 (56.7291, 41.0919)	47.779 (55.2304, 40.0635)	48.7708 (56.4311, 41.4345)	49.0222 (57.4878, 41.2195)	48.1251 (55.8614, 40.6546)	48.0616 (55.3207, 41.0826)
36	Oestroidea	42.4538 (53.384, 29.6454)	37.4747 (48.8761, 25.1366)	41.5834 (51.3, 30.2483)	43.8944 (55.3986, 29.7584)	38.4561 (48.6491, 23.15)	40.5693 (50.2828, 28.4536)
37		29.9744 (43.8728, 14.9316)	25.8197 (38.9425, 11.1851)	32.2801 (44.8406, 19.1293)	31.2649 (46.4283, 14.6572)	26.7526 (39.443, 12.3544)	29.6672 (42.228, 16.3938)
38		42.3043 (49.0759, 35.7578)	40.4896 (47.1478, 33.4808)	41.558 (48.0874, 35.0866)	42.4745 (49.2803, 35.8447)	41.1596 (47.9065, 34.6534)	41.0162 (47.2076, 34.7468)
39		56.2968 (72.034, 33.4165)	52.0757 (68.6714, 30.4499)	54.2845 (68.0322, 36.6271)	58.3718 (74.0658, 37.4291)	53.1886 (69.4261, 32.5562)	52.6837 (68.3015, 31.1419)

however, several changes in the results generated from the concatenated data set. Node 27 increased from 106 my with gene based partitioning to 114 my, a result more consistent with other estimates. Conversely, the 30 my age estimate produced for node 30 when using codon based partitioning was at least 20 my younger than the age estimate produced using other data sets and methods.

In general, using a codon based partitioning scheme had a small positive effect on node ESS's in 3rd codon position excluded data sets and a greater impact on 3rd codon position included data sets. More notably, use of fewer partitions greatly increased the ESS of the tree posterior for analyses which used 3rd codon position excluded data under the GTR model.

Discussion

Mitochondrial vs. nuclear gene data sets

In our analyses, both mitochondrial and nuclear gene data sets gave remarkably similar results for the vast majority of the nodes in our tree despite notable differences in sequence evolution. Nonetheless, the two data sets can not be said to perform equally well. Several nodes proved to be far more difficult to estimate with mitochondrial gene data than with nuclear genes, and when conflicts existed between mitochondrial and nuclear clade age estimates examination of the trace data usually showed the results from nuclear genes were less noisy.

A priori, concatenation of the two data sets could produce three possible outcomes: an age estimate that represents an intermediate point between the data sets due to near equal support being present in both sets, an age estimate independent of the two estimates (either higher or lower than either set alone) due to the increased volume of

data improving estimation, or, most desirably, support for one result strongly over the other due to consistent signal in one source and weak support in the other. In our results, we most often saw the first case, where the concatenated data set produced an age between the mitochondrial and nuclear ages. Thus, while the concatenated result produces an estimate consistent with the total evidence, it does not serve to resolve disputes between data sources or function better than either data set alone.

Third codon positions

Ideally, 3rd codon positions should be capable of producing divergence time estimates as well as first or second codon positions if they are modeled properly. Furthermore, inclusion of 3rd codon position data could increase the efficacy of divergence time estimation on more recent divergences as their exclusion results in the *de facto* elimination of fast evolving sites which are likely to contain information on the shortest internodes and most recent events. This is, however, an optimistic expectation. As 3rd codon positions are subject to significant amounts of homoplasy over longer evolutionary distances, they are likely to introduce noise into the data set and reduce resolution of more ancient nodes where multiple substitutions are more common. Due to the increased homplasy, we also find that 3rd codon positions were more affected by substitution biases leading to increased divergence in base composition.

Our results showed that third codon position data did not add appreciably to the value of our calculations when data were partitioned by gene. While estimates including third codon positions were frequently very close to their third codon excluded counterparts, ESS's were reduced indicating they have increased the complexity of the calculation for no practical benefit. When data was partitioned by codon position rather

than gene, we found that third codon positions had a noticeable negative impact on our ability to infer ages. Interestingly, there was no obvious time depth dependent effect of third codon inclusion on either inferred age or ESS in the span of time covered by our tree as might have been predicted by previous studies (Phillips 2009).

Model complexity

The issue of model fit vs. overparameterization/overfitting is one familiar to molecular evolution researchers (Rannala 2002; Sullivan, Joyce 2005). While an appropriate complex model will almost always fit the data better than a simpler model, the increased fit can come at significant computational cost and the introduction of more parameters to estimate increases the likelihood of errors creeping into the results. Our alignments represent a fairly complex data set with a total of 23 genes evolving in two distinct genomes over a 200my time period. Thus we tested the efficacy of the more complex GTR model vs. the popular but simpler HKY model to investigate what impacts an improved model would have.

We found that the more complex GTR model performed no better on our data set than the simpler HKY model when our data set was partitioned either by gene or by codon position. Consistent with its greatly increased complexity, the GTR model produced lower ESS's for the same nodes; however, nodes for which both data sets (HKY and GTR) produced acceptable ESS's produced nearly congruent results. This indicated that analyses using the GTR model would require many more generations to sample the data than those using the HKY model, yet the GTR model did not produce an improved estimate in most cases.

Partitioning schema

Partitioning of data sets allows us to specify *a priori* what regions of a data set are known to be "different" from other regions and estimate model parameters independently for each these partitions. Two naïve approaches for partitioning data sets naturally recommend themselves to the researcher: creating a separate partition for each gene and creating a separate partition for each codon position. Combination of the two methods is also an option, although a great number of small partitions are required. Moreover, the limited information available in each partition would likely have negative impacts on parameter estimation (Rannala 2002). Between the two partitioning options, which to choose depends heavily on how the researcher visualizes the evolution of the genes under study. For multiple genes evolving at heterogeneous rates, consistent with our nuclear gene data set, an assumption of higher variability between genes than between first and second codon positions within the same gene would likely be reasonable. For a set of genes evolving at a roughly similar rate or characterized by skewed base composition between first and second codon positions, a situation consistent with our mitochondrial gene data set, concatenating the genes and creating separate partitions based solely on codon position would be the obvious choice. When a highly heterogenous data set such as the one investigated in this study presents itself, however, the choice of how to properly partition the data is not an obvious one.

Our results showed little difference between codon and gene partitioning when third codon positions were excluded. For mitochondrial genes and concatenated data sets using an HKY model, by codon partitioning gave slightly superior results to by gene partitioning. When using the GTR model, the improvement in mitochondrial gene estimates by using codon based partitioning over gene based partitioning was more

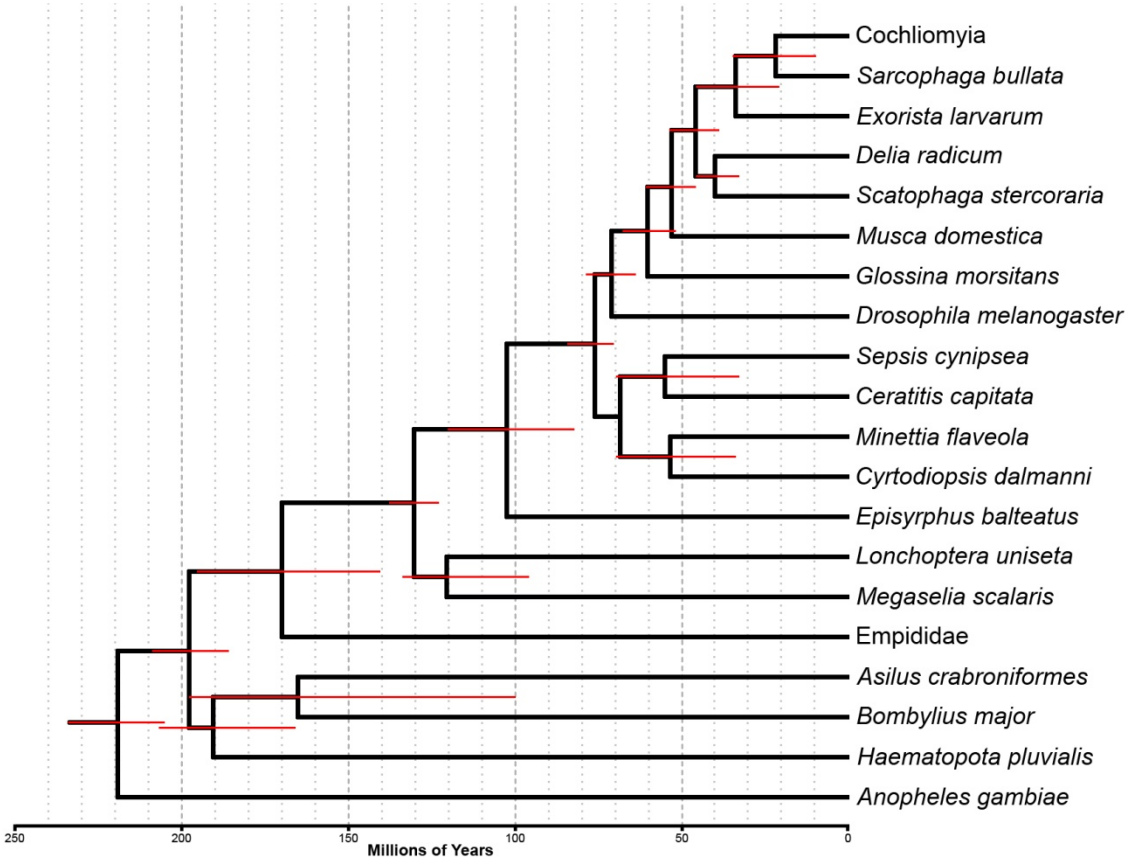
obvious. While this difference may have been due to the great decrease in the number of parameters requiring estimation in codon versus gene partitions (2 models vs. 13 models), it is notable that the nuclear gene data showed no such improvement in ESS's when compared even though a similar reduction in parameters was achieved (2 models vs. 10 models).

Implications for dipteran phylogeny

The convergence of our clade age estimates across multiple data sources and methodologies indicates highly robust support for these dates throughout the majority of nodes covered in our tree. Comparison of our age estimates to those arrived at for the same nodes in Wiegmann *et al.* (2011) shows only a relatively small disagreement. Our calibrated age for the culicomorphan/brachyceran divergence (node 21) is nearly 20my younger than the estimate arrived at in Wiegmann *et al.* (2011) (Fig 3.4). The same is true for our calibration for the age of the cyclorrhaphan crown group (node 26). The brachyceran and schizophoran calibration points (nodes 22 and 29), however, are within approximately five my of the ages estimated in Wiegmann *et al.* (2011). While two of the three deepest calibration points in our tree were arbitrarily constrained to possibly exclude a portion of the likely age distribution, a similar criticism could be applied to any other assigned prior. Ultimately, the true distribution of possible ages can not be known with any certainty and an arbitrary distribution must be chosen. Furthermore, as these two calibration points were isolated from the majority of taxa included in the study, their influence on clade age estimates within the orthorrhaphous Brachycera or our target group, the Schizophora, was likely to be minimal.

Figure 3.4. Chronogram

Horizontal scale indicates node age in millions of years. Nodes are placed at the median age estimate from the nuclear gene 3rd codon position excluded analysis, HKY model. Red bars indicate bounds of 95% confidence interval.



Within the non-schizophoran nodes of our tree, our data set produced ages congruent with those arrived at in prior studies (Wiegmann et al. 2003; Wiegmann et al. 2011). We placed the divergence of the Tabanamorpha from the Asiloidea (node 23) at approximately 192 mya. Inference of the age of the Asiloidea (node 24) posed particular challenges when using mitochondrial gene data; however, nuclear gene data alone consistently produced an age of approximately 165 my for this clade. The age of the Eremoneura crown group (node 25) is consistently estimated to be approximately 172 my, although when third codon position data are included, mitochondrial gene data alone produced median age estimates as young as 157 my for this clade. Considering the generally negative effects we observed from adding third codon position data to our analyses and the agreement of the concatenated data set with the 3rd codon excluded results, the 172 my age for the Eremoneura should be considered the more robust estimate. The divergence time of the crown Platypezoidea (node 27) showed some discrepancy between mitochondrial and nuclear gene estimates, typically being resolved to between approximately 95 mya and 120 mya depending on data source. Examination of the traces for both data sets revealed distributions skewed towards older age estimates, with the width of the mitochondrial distribution being significantly wider. The true age of this node likely lies somewhere between the 104 my age estimate derived from the concatenated data set and the 120 my estimate derived from nuclear gene data alone; however, it is also possible that the 125-135 my constraint placed on the adjacent cyclorrhaphan node (node 26) confined our ability to estimate of this node and that the true age is even older. For the final non-schizophoran node we investigated, we found the syrphids to have last shared an ancestor with the schizophoran flies roughly 100 my

(node 28). Once again, a small discrepancy exists between nuclear and mitochondrial gene data sets and the concatenated data set was in close agreement with the nuclear gene estimate.

Within the "acalyptrate" schizophoran flies (nodes 30-32, 39), the relationships between taxa are not firmly established (Wiegmann et al. 2011), and our taxon sampling within this group was not comprehensive. Nonetheless, the tree we used represents our current best estimate of schizophoran relationships and our results can be viewed as the foundation for more in-depth work on this clade. We found strong agreement between mitochondrial and nuclear data sets for all nodes in this group except node 39 where an approximately 7 my discrepancy was observed. Investigation of the trace data for node 39, the *Minettia flaveola*/*Cyrtodiopsis dalmanni* divergence, suggests that the older 60 my age estimate derived from mitochondrial gene data may be the more accurate estimate in this case. Clade age estimates estimates for all major schizophoran lineages, including the Calyptratae (node 33) lay in the range of 55-72 my. This result is consistent with the hypothesis of an explosive radiation at the base of the schizophoran tree (Blagoderov, Grimaldi, Fraser 2007).

For the Calyptratae, internal species relationships are better supported and there are no major discrepancies between nuclear and mitochondrial clade age estimates. This instills confidence that our estimates provide a meaningful first molecular framework for divergence times of major calyptrate clades. We calculate the age of the calyptrate crown group (node 33) to be approximately 59 my. The paraphyletic clade containing both "Muscoidea" and Oestroidea (node 34) appeared 52 mya, and the divergence of Anthomyiidae from Scatophagidae (node 38) occurred approximately 41 mya. The

Oestroidea crown group (node 36) appeared approximately 36 mya, and Calliphoridae diverged from Sarcophagidae (node 37) approximately 25 mya. It should be noted that node 38 (the Anthomyiidae/Scatophagidae split) was one of our calibration points with 80% of the distribution contained in the interval from 42-52 mya, but the age estimate is consistently younger by several million years. This is the only calibrated node for which the age estimate diverged notably from the mean of our assigned age distribution, producing an age younger than our fossil calibration point. Therefore, there may be a tendency to underestimate the age of this node and possibly other nodes within the calyptrates in our analysis.

Conclusion

Overall, we see highly congruent results between different data sources, models, and partitioning schemes. These results indicate highly robust support for clade age estimates arrived at under a variety of analytic regimens. Considering the degree of convergence between these estimates, we suggest that optimizing computational time, fossil calibrations, and sampling efficiency should take precedence over optimization of model fit and fine tuning of data preparation when calculating clade ages of similar time depths to those observed within the Diptera. Towards this end, we formulate several specific suggestions for researchers seeking to optimize their results.

Recommendations for efficient research design

First, we suggest that nuclear encoded genes should be preferred over mitochondrial genes in the time range of 30-220 my if a choice must be made; however, comparison of the age estimates derived from both sources can be informative if the data and computational resources are available. Second, 3rd codon positions should be

excluded from the data set when investigating divergences in timeframes similar to the one we investigated. While their presence had little observable effect on clade ages in our data set, they did have a negative impact on ESS indicating an increased potential for misestimation. Third, unless there is a strong reason to prefer a more complex model, we suggest using a less parameter rich model such as HKY. We found that estimations using an HKY model were nearly identical to those produced under the more complex and better fitting GTR model but overall sampling efficiency was greatly improved under the HKY model. Lastly, as there was little effect on inferred age under different partitioning schemas, we suggest partitioning data by whichever method seems more appropriate or convenient unless using the GTR model. With the more complex GTR model, reducing the partition count by using a codon position based partitioning scheme greatly improved sampling efficiency.

Comparison to previous work

Our results present an interesting contrast to those of Phillips (2009), which dealt comprehensively with similar issues of model selection and data preparation in a manner complementary to our own. Phillips' results suggest that using a less complex model, such as HKY rather than GTR, or increasing homoplasy in the data, such as by inclusion of third codon position data, would lead to time depth dependent misestimation of clade ages. This predicted result was not obviously visible in our analyses; however, our data set displayed several important differences from Phillips' test data set which may contribute to this discrepancy. First, the deeper nodes in our tree where we would expect to see the largest impact of branch length misestimation are typically calibrated nodes. As by design our calibration points were tightly constrained, these nodes and the handful

of deep level nodes which were not calibrated had little flexibility in their placement. As noted in our methods, relaxing the constraints on our calibration points led to a shift in estimated divergence times, although no comprehensive effort was made on our part to explore the degree of misestimation across data sets. It is notable, however, that our most variable age estimates, the Asiloidea (node 24) and Platypezoidea (node 27), are deeper nodes not constrained by younger calibration points or shallower nodes. We attribute these difficulties to poor sampling (ESS) of the nodes in question. In the case of Asiloidea, this is possibly due to base composition biases within the mitochondrial genome. Alternatively or in addition, one or both of these nodes may be varying due to branch length misestimation. If such is the case, it seems most likely that the differences between mitochondrial and nuclear gene evolution are the more important factors at work, as third codon position inclusion and use of the HKY rather than the GTR model had little impact on inferred ages.

A second consideration is that our data set concentrates on a time span of approximately 220 my, which is notably shallower than the 420 my covered by Phillips' vertebrate data set. Severe biases may not begin to manifest within the time frame covered by the Dipteran radiation when analyzed with the fossil calibration points we chose.

Lastly, while our methods were analogous to those of Phillips', they were designed to compare common "use case" scenarios rather than to tease apart one specific cause of clade age misestimation. It is possible that our data preparations do not vary sufficiently to highlight time depth dependent effects.

We have found that both mitochondrial encoded and nuclear encoded genes produce largely congruent age estimates for most Dipteran clades groups. The cases where estimates diverge between data sets indicate that biases present in the data can locally affect the age estimates of select nodes without adverse impact on the remainder of the tree. Our study leaves unresolved the question of what the specific causes of these incongruencies are. Whether they are the result of "rogue taxa" creating a local misestimation of the node they are directly attached to, alterations in the substitution patterns of a particular branch of the tree, or unaccounted for systematic biases in one or both data sets that manifests as misestimation of a certain class of nodes is a question that future research may answer. As molecular divergence time estimation has become a ubiquitous part of modern phylogenetic analysis, answers to these questions and methods of limiting their impact would be welcomed by evolutionary biologists.

Acknowledgements

The authors would like to thank the Wayne State High Performance Computing Grid for computational resources used in this study and the members of the Assembling the Tree of Life: Diptera project for contributing sequence data and resources. This project was funded through NSF award EF-0334948. JC was recipient of several Thomas C. Rumble fellowships by Wayne State University.

CHAPTER 4 “DISCUSSION”

Simulation studies and empirical test sets

When testing phylogenetic methods, there are two main approaches to data set design. Simulated data sets which are artificially “evolved” with specified constraints represent one powerful tool for teasing apart phylogenetic methods. These data sets allow the researcher to specify all aspects of sequence evolution, including branch lengths, substitution patterns, and tree topologies. Simulated data sets are powerful tools for answering very specific questions of algorithm performance as all variables can be controlled and only a single parameter changed between simulations. Likewise, as these are artificially generated, all parameters are known and thus the truth of a result can be directly determined from the models used to create it.

An alternative approach is to use empirical data from real world data sets. These data sets do not necessarily fit any known evolutionary models and have been evolving under totally unknown constraints, usually for millions of years. In empirical data, substitution patterns and selection constraints may have shifted multiple times over the course of evolution, population bottlenecks may have resulted in local alterations to the rate of substitution fixation, evolutionary novelties may have resulted in selective sweeps, or external factors such as disease, predation, or a changing environment may have increased selection pressure on certain taxa. In general, empirical data reflects the full range of evolutionary scenarios that impact evolution at both the macro and micro level.

Empirical data does not lend itself as well to testing narrow questions as its evolution was not controlled. The substitution processes which created the real world data set are unknown and must be inferred from the data, unique replicate data is not

available, and the truth of a result can not be strictly quantified. Thus, for experiments which address the behavior of tree reconstruction under specific circumstances, simulation data is superior because it allows the researcher to fix all parameters irrelevant to the question at hand and carefully control the parameter of interest.

However, simulated sequence data, while constantly being improved, is still biologically unrealistic. Artificially evolved sequences are the embodiment of the biases of the algorithm and parameters used to generate them and are thus constrained in a way that empirical data are not. Problems with the simulation of more complex evolutionary processes such as the poorly characterized insertion/deletion process and maintenance of locally conserved sequence regions are still very common (Strope et al. 2009), and unknown or difficult to quantify processes are likely not represented at all. Methods for simulating data sets are improving, recently with particular attention being paid to the simulation of whole genome sequences (Earl et al. 2011), however they are currently still limited. Empirical data sets do not share these problems. Since empirical data are not evolved under known models, all of the complexity of natural evolution can be present in the data. Furthermore, all parameters for the analysis of simulated data must be estimated and inferred from the existing sequences. As working backwards from existing data to discover the processes which gave rise to them is the usual method for phylogenetic inference, empirical data is more suitable for direct comparison of methods.

In my analyses, I chose to use empirical data rather than simulated data. The questions I was asking about the phylogenetic utility of mitochondrial and nuclear genes did not lend themselves to the use of simulated data because the question was not narrowly defined in terms of controllable sequence evolution parameters. As I intended

to explore whether mitochondrial genes or nuclear genes offered superior phylogenetic utility, only empirical data sets could properly reflect the complexity of the issue in a way that would prove instructive to future researchers. The performance of "mitochondrial-like" or "nuclear-like" simulated sequences with all of the limitations and simplifications involved in simulation is not as informative or compelling as the performance of actual real world data sets.

Diptera as an evolutionary test data set

The AToL: Diptera project was established to provide a comprehensive re-examination of dipteran relationships. In addition to re-scoring morphological data matrices, a large volume of DNA sequence data was gathered with an eye towards thorough and even taxonomic sampling. The DNA sequence data was gathered in two stages. The Tier 1 group was sequenced for 14 nuclear genes and complete mitochondrial genomes. 42 species representing major infraorders and families were sequenced in this manner. The deep sequencing of the Tier 1 taxa was intended to provide a high quality backbone phylogeny of Diptera. The Tier 2 group included 202 species, sequenced only for 5 nuclear genes. These more lightly sequenced taxa were selected as exemplars to resolve family and genus level relationships as well as contribute to the backbone phylogeny arrived at with the Tier 1 taxa sequences.

The data set of the AToL: Diptera project provides a convenient and useful real world data set for the testing of the phylogenetic utility of mitochondrial and nuclear gene sources. The Diptera present a complex and non-trivial example of evolutionary complexity. Tremendous morphological and lifestyle diversity are present within the clade; a relatively steady pace of diversification has been maintained with family ages

ranging from approximately 240 myo to 22 myo; and there are several periods of explosive radiations which complicate phylogenetic inference. Nonetheless, many major clades within the Diptera are morphologically distinct and non-controversial thus allowing the "truth" of any inferred tree to be evaluated and a reasonably intact fossil history provides us with calibration points and guidelines for the evaluation of divergence time estimates.

The availability of such a large data set which contains both mitochondrial and nuclear gene data is a boon to evolutionary biologists studying phylogenetic methods. The variety of clade ages, the challenging to infer topologies at rapid radiations events, and the presence of well resolved clades which serve as known benchmarks all contribute to its power as a test data set. The AToL: Diptera data set provides a useful test set for the study of an assortment of phylogenetic questions and methods.

Concatenation of mitochondrial and nuclear gene data improves clade recovery

My results showed a positive effect from the addition of complete mitochondrial genomes to sampled nuclear genes. This effect went beyond the mere strengthening of branch support values that may be expected due to increased volume of sequence data. Rather, I saw branches where nuclear gene data alone is insufficient to resolve a relationship, however concatenated mitochondrial and nuclear gene sequence data resolves it with high support. Furthermore, when I observed topological discrepancies between mitochondrial and nuclear gene trees, concatenation of the data sets typically resolved the dispute in favor of the more historically favored topology. While this typically resulted in favoring the nuclear gene tree topology over the mitochondrial topology, branch support for conflicting nodes was robust in trees derived from

concatenated data sets indicating no obvious deleterious effect resulting from the inclusion of the conflicting mitochondrial data.

These results are exciting for researchers in molecular phylogenetics. While nuclear gene data proved to be a more reliable estimator of phylogenetic relatedness than did mitochondrial gene data, the addition of mitochondrial gene sequences to nuclear gene sequences provided an overall positive effect with no noticeable downsides. For targeted phylogenetic studies in groups where nuclear sequences may be particularly difficult to obtain due to extreme sequence divergence, allelic differences, gene duplications, or other confounding effects, the addition of relatively easily obtained mitochondrial gene sequence data to whatever nuclear gene data can be obtained can provide additional robustness to the results. These results may also be encouraging to researchers performing phylogenetic studies of very species-rich groups which demand extensive taxon sampling. Fewer of the relatively difficult to amplify nuclear genes can be sampled and replaced with easily obtained longer mitochondrial gene sequences with little risk of biasing resulting trees.

Mitochondrial and nuclear gene data are not equivalent estimators of divergence time

For many clades, I found that divergence time estimates produced from mitochondrial genes were similar to those produced by nuclear genes when analyzed with the BEAST software (Drummond, Rambaut 2007; Rambaut, Drummond 2007). However, notable exceptions were found which indicated inferior performance of mitochondrial genes on several nodes. These results indicate that previous studies which used only mitochondrial genes as estimators of divergence time should be viewed with

some skepticism. While I found agreement between the two sets of estimates for the majority of nodes, the exceptions were in some cases quite extreme. Furthermore, the majority of published divergence time estimates do not include ESS or equivalent metrics, so identifying which specific nodes may be problematic and which are robustly resolved is often impossible. When both independent nuclear and mitochondrial age estimates are available for a group, I suggest that the nuclear gene derived estimates be preferred.

Influence of 3rd codon positions on divergence time estimates

I found that inclusion of 3rd codon positions is generally not desirable in divergence time estimation at the time depth we studied. While estimates from data sets which included 3rd codon position data were not noticeably biased, they increased computational complexity and did not result in an increase in clade age resolution.

At first glance, these results appear to stand in contrast to the recent results which suggest 3rd codon site inclusion as essential to accurate age estimation (Yang, Yoder 2003). My methods and my data set differ notably from those of Yang and Yoder, however. In their study, only two mitochondrial genes were used rather than the 23 genes I studied. Their trees covered a time span of only 90 my with their group of interest being less than 10 my old while my results covered a time span of over 200 my with my groups of interest being approximately 100 my and younger. Lastly, they do not consider the case of 3rd codon position excluded data sets and instead compare only each codon position in isolation to all 3 positions. These differences suggest several possible explanations for why my results differ. First, they compared single codon position data sets from only two genes. As parameter estimation is improved on larger data sets

(Rannala 2002), it is likely that my 13 gene combined mitochondrial gene data set provides a better overall estimate for codon position evolution. Second, as 3rd codon positions tend to have rapid rates of substitution, homoplasy will increase over time. While it may be the case that mitochondrial 3rd codon positions are informative on lemur divergences of under 10 my, they may not hold sufficient signal to resolve my older cyclorhapphan relationships. Finally, they did not test combined 1st and 2nd codon positions, thus I do not know whether 1st and 2nd codon positions would have produced similar results to the results they obtained from all 3 codon positions as I saw in my analysis.

I suggest that 3rd codon positions be excluded from divergence time analyses at time depths of approximately 40 my and older. I saw some small evidence that 3rd codon positions may have had some influence on divergence time estimates for the most recent nodes in my tree (<40 my), however the change was still very small (~ 3 my change to median) and it was not clear whether this reflected an increase in accuracy or a misleading bias as the true clade ages are unknown. The more notable effect 3rd codon positions had was on ESS values of clade ages. These values suffered visibly from the addition of 3rd codon positions and lower ESS is clearly linked to reduced accuracy. As such, I see little reason to use these sites for older time depths.

For divergence time estimation, simpler is better

I found that my efforts to increase the size of my data set or to model it with more precise models did not result in improved accuracy of divergence time estimates. Concatenation of mitochondrial and nuclear gene data did not produce divergence time estimates that were visibly improved relative to using a single data source. Using the

more complex GTR model rather than the HKY model also failed to produce any improvement in clade age. Lastly, a more parameter rich "per gene" partitioning scheme did not produce improvements over the simpler "per codon" partitions. Instead, each one of these measures caused ESS of the samples to fall and therefore reduced the ability of the BEAST program (Drummond, Rambaut 2007; Rambaut, Drummond 2007) to explore clade age parameters.

I suggest that in this case, improving model fit by increased model complexity has a performance penalty that is not commensurate with any improvements it may offer in terms of accuracy. While the ESS could be improved by exploring parameter space for a longer period of time, there is no evidence that the analyses with more complex parameterization produced any benefit to the resolution of clade ages for those nodes which had sufficiently large ESS to consider them well resolved. This represents a clear example of over-parameterization of a phylogenetic question.

Implications for the resolution of the Dipteran phylogeny

My results verified many of the well established clades of Diptera. I successfully recover Eremoneura, Brachycera, Cyclorrhapha, Schizophora, Calyptratae, and Oestroidea with robust support. I also recovered a monophyletic Asiloidea and the two sampled bibionomorphs were monophyletic as well. The recovery of these benchmark clades suggests that my methods and data set was recovering the tree accurately.

More interestingly, I confirmed the sister group of both Schizophora and Calyptratae. The relationships of the "lower cyclorrhapan" groups to the Schizophora have long been a topic of debate (Yeates, Wiegmann 1999). My results show strong support for a Syrphoidea + Schizophora clade, in agreement with recent results from

other phylogenetic studies (Wiegmann et al. 2011). Consistent with recent findings, I also recovered Drosophilidae (representing Ephydroidea) as sister to the Calyptratae (Hwang et al. 2001; Cameron et al. 2007; Wiegmann et al. 2011). The calyptrate flies have long been recognized as a distinct monophyletic clade within the Schizophora, however their relationship to other schizophoran flies was the subject of much speculation. Furthermore, as the ephydroid fly *Drosophila melanogaster* is possibly the most popular animal model system, locating the sister group of the Ephydroidea places all of the accumulated data on *D. melanogaster* in its proper context for scientists interested in comparative evolution. This finding is thus of great benefit to both dipterologists and evolutionary biologists in general.

Unfortunately, even the large Dipteran data set produced by AToL: Diptera was not sufficient to resolve the relationships of the remaining schizophoran taxa. Neither my results nor those of Wiegmann *et al.* (2011) resolved these relationships with high confidence. My results for relationships within this clade do not agree with those of Wiegmann *et al.* (2011), however neither study produced strong support in favor of any single topology. These relationships have proven problematic to resolve in the past due to the likely rapid radiation of basal members of the clade (Wiegmann et al. 2003; Wiegmann et al. 2011), thus this result is not surprising. It was hoped, however, that the scale of the AToL: Diptera sequencing effort would be sufficient to resolve these clades.

Perhaps most interestingly, my results and those of Wiegmann *et al.* (2011) suggest that the relationships of basal brachyceran groups must be reevaluated. Prior to these two recent molecular studies, an infraorder dubbed Muscomorpha, comprised of Asilomorpha and Eremoneura, was one of the most accepted feature of basal brachyceran

relationships (Woodley 1989b; Yeates, Wiegmann 1999) with the remaining brachyceran infraorders largely unresolved. Wiegmann *et al.* (2011) produced a tree which resurrected a largely disregarded grouping named Orthorrhapha which joined all non-eremoneuran brachycerans into a monophyletic clade that formed the eremoneuran sister group. My results support neither Muscomorpha nor Orthorrhapha as the correct topology of basal brachyceran groups. Instead, we recover the horse flies, Tabanamorpha, as the most basal brachyceran group and a clade comprised of Asilomorpha and Stratiomyomorpha as the Eremoneuran sister group. These competing brachycean topologies are sure to be the subject of targeted phylogenetic efforts in the near future.

First divergence time estimates for major calyptrate families

The divergence times of the calyptrate groups are not known, with only a few scattered fossils, mostly of ancient stem groups (Grimaldi, Engel 2005) (T. Pape, personal communication). Thus, I used molecular divergence time estimation to produce the first estimates for these clades.

My results showed the crown clade comprised of the paraphyletic Muscoidea and the Oestroidea to be approximately 53 myo. The Anthomyiidae clade of leaf miners and plant parasites and the Scatophagidae clade of dung and detritus feeders as well as plant parasites diverged approximately 47 mya and last shared an ancestor 40 mya. The Oestroidea crown group arose 35 mya, and the mammalian parasite bot and flesh flies diverged from each other 22 mya.

Future directions

My work suggests several possible avenues for future exploration, both to expand on my methodological conclusions and to further investigate the less well resolved regions of the dipteran tree.

While my evidence in favor of the inclusion of mitochondrial genomes along with nuclear genes is compelling from a procedural standpoint, the underlying question of phylogenetic signal distribution between data partitions has not been addressed. It is clear that trees derived from mitochondrial genes alone are not as well resolved or as accurate as those derived from nuclear genes, thus the mitochondrial genes must contain conflicting or extremely weak signal. At what point these signals would drown out or merely fail to contribute to the nuclear gene signal is unknown. A comprehensive analysis of varying amounts of nuclear and mitochondrial gene data is necessary to detect at what point nuclear gene derived signal is not strong enough to overcome the mitochondrial gene signal for conflicting topologies. Furthermore, a subset of mitochondrial genes rather than the entire protein coding content may be optimal. This was not tested in my analyses, however it is a logical extension of my work as phylogenetic signal is likely not homogenous across the mitochondrial genome. Lastly I did not investigate data sets which included mitochondrial rRNA or tRNA sequences as my focus was partially on the effect codon positions on branch recovery. A more thorough investigation of how this additional mitochondrial data may impact branch recovery would be helpful for future research. It is quite possible that these additional sequences would further increase the value of adding mitogenome data to an analysis.

My divergence time analysis also suggests several interesting questions with far reaching ramifications. I provide evidence that divergence date estimates derived from mitochondrial and nuclear genomes are not equivalent within the Diptera over the spread of ages covered by my tree. It is unknown whether this effect is Diptera specific or whether it is generalizable. Likewise, very different behavior may be observed in younger or older clades than those I investigated in the Diptera. These questions must be answered as a sizable body of mitochondrial gene derived clade ages exists and my results call their accuracy into question. I also produced results which suggest that complex evolutionary models were responsible for over-parameterization of the problem space and resulted in degraded resolution at some nodes. It is not known what impact increased parameterization would have on other data sets which differ in size or composition when compared to ours. While I believe that my results are instructive for model selection, I cannot discount the possibility that more complex models and partitioning schemes may be crucial to resolving some clade ages.

I provided a robust tree of dipteran relationships including new hypotheses on basal brachyceran relationships, and updated my understanding of which parts of the dipteran tree I can take for granted and which clades I must still view as tentative. My results only serve as a starting point, however, and must be verified by narrowly targeted work. Comprehensive taxon sampling in the basal Brachycera and the non-calyptrate Schizophoran was not a priority in my analyses, thus it is possible that my results may be artifacts of insufficient sampling. Targeted sequencing of select basal brachyceran and "acalyptrate" sequences may improve resolution in these areas of the tree and resolve the questions I raised.

APPENDIX A “ARBIVORE.PL”

```
#This program reads in newick formatted tree files and determines statistics associated with nodes
#The contents of the clades it looks for can be edited
#This particular implementation of the script reads in an external file containing divergence
#dates and outputs a tab delimited spreadsheet which contains information on which clades were
#recovered by which data sets and what the branch support assigned to that node was.

use strict;
use warnings;

use Bio::Phylo::Factory;
use Bio::Phylo::IO;

my $factory = Bio::Phylo::Factory ->new;

#Dates table must be created from divergence time data.
#Format of file: Node#\tnuclear median (min, max)\tmito median (min, max)\tconcat median (min/max)\n
my $dates_file = "dates_table.xls";
our $dates_hash = parse_date_file ($dates_file);
#0 for nuc, 1 for mito, 2 for concat
our $date_index = "0";

my $clade_hash = create_clade_hash();


open (my $out_fh, ">", "clade_stats.xls") or die $!;

print $out_fh "File\tMethod\tSource\tType\tSites";

foreach my $clade (@clade_order) {
```

```
        print $out_fh "\t$clade support\t$clade branch length\t$clade avg depth\t$clade
min depth\t$clade max depth";

}
print $out_fh "\n";

my @trees = <*.tre>;
push (@trees, <*.nwk>);

foreach my $treefile (@trees) {

   my $params_ref = parse_file_name ($treefile);
   my $params = join ("!", $treefile, @$params_ref);
   $params =~ s/\!/\t/gi;

   print $out_fh "$params";
   print"$params";
   my $type = $params_ref -> [3];
   print "\nType: $type\n";

   my $tree_string = parse_file ($treefile);


   my $forest = Bio::Phylo::IO->parse(
    -format => 'newick',
    -file => $treefile
   );


   my $tree = $forest -> first;
   my @internals = @{$tree -> get_internals};
   foreach my $internal (@internals) {
      my $name = $internal -> get_name;
      if ($name  && $name <= 100){
         $internal -> set_score ($name);
      }
```

```
    }


    foreach my $key (@clade_order) {
           print "clade = $key\n";

           #Added for divergence time changes
           unless (exists $dates_hash -> {$key} ) {
                  next;
           }


       my $ancestor_node = identify_ancestor ($tree, $clade_hash -> {$key}, $type);
       if ($ancestor_node) {
          print "$key found!\n";

          #Added $key to process ancestor for divergence time stuff
          process_ancestor_node ($ancestor_node, $key);
       }
       else {
          print "$key not found!\n";
          print $out_fh "\t\t\t\t\t";

       }
    }
    print $out_fh "\n";
}




sub parse_file {
    my $file = shift;
    open (my $tree_fh, "<", $file) or die;
    my $return_string;
    while (my $line = <$tree_fh>) {
```

```
      chomp $line;
      $return_string .= $line;
   }

   return $return_string;
}




sub identify_ancestor {
   my $tree = shift;
   my $clade_ref = shift;
   my $type = "nuc";


   my %trimmed_clade_hash = %$clade_ref;

   foreach my $key (keys %$clade_ref) {
      if (($clade_ref -> {$key} eq "both") || ($clade_ref -> {$key} eq $type) ) {
         next;
      }
      else {
         delete $trimmed_clade_hash {$key};
         print "deleted $key\n";
      }
   }

   my @internals = @{$tree -> get_internals};

   NODE:foreach my $node (@internals) {
      my @terminals = @{$node -> get_terminals};
      if (@terminals == keys %trimmed_clade_hash) {
         my $number = @terminals;
         print "testing $node\tright number of taxa\t$number\n";
         foreach my $taxa (@terminals) {
```

```
            my $name;
            if ($taxa -> get_name) {
               $name = $taxa -> get_name;
            }
            else {
               $name = "unknown";
            }
            if (!exists $trimmed_clade_hash{$taxa -> get_name} ) {
               "$name does not exist!\n";
               next NODE;
            }
         }
         return $node;
      }
      else {
         my $number = @terminals;
         #print "skipping $node\twrong taxa count\t$number\n";
         next NODE;
      }
   }




}



sub create_clade_hash {

   my %brachycera =  (  "Acrabronif",  "both",
                        "Bmajor",      "both",
                        "Ccapitata",   "both",
                        "Cdalmanni",   "both",
                        "Cochliomyi", "both",
                        "Dradicum",    "both",
```

```
                    "Dmelanogas",  "both",
                    "Ebalteatus",  "both",
                    "Elarvarum",   "both",
                    "Empid",       "both",
#                    "Eangustrif",  "both",
                    "Gmorsitans",  "both",
                    "Hillucens",   "both",
                    "Hpluvialis",  "both",
                    "Luniseta",    "both",
                    "Mdomestica",  "both",
                    "Mflaveola",   "both",
                    "Mscalaris",   "both",
                    "Sbullata",    "both",
                    "Scynipsea",   "both",
                    "Sstercorar",  "both"
                   );
#Note: Incompatible with Orthorrhapha
my %muscomorpha = (  "Acrabronif",  "both",
                    "Bmajor",      "both",
                    "Ccapitata",   "both",
                    "Cdalmanni",   "both",
                    "Cochliomyi", "both",
                    "Dradicum",    "both",
                    "Dmelanogas",  "both",
                    "Ebalteatus",  "both",
                    "Elarvarum",   "both",
                    "Empid",       "both",
                    "Gmorsitans",  "both",
                    "Luniseta",    "both",
                    "Mdomestica",  "both",
                    "Mflaveola",   "both",
                    "Mscalaris",   "both",
                    "Sbullata",    "both",
                    "Scynipsea",   "both",
                    "Sstercorar",  "both"
                   );
```

```
#Note: Incompatable with Muscomorpha
my %ortho        =  ("Acrabronif",  "both",
                     "Bmajor",      "both",
                     "Hillucens",   "both",
                     "Hpluvialis",  "both"
                    );

#Note: Incompatable with Muscomorpha
my %ortho1       =  ("Acrabronif",  "both",
                     "Bmajor",      "both",
                     "Hillucens",   "both"
                    );


my %asiloidea = (   "Acrabronif",  "both",
                     "Bmajor",      "both",
                    );

my %eremoneura = (   "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Cochliomyi", "both",
                     "Dradicum",    "both",
                     "Dmelanogas",  "both",
                     "Ebalteatus",  "both",
                     "Elarvarum",   "both",
                     "Empid",       "both",
                     "Gmorsitans",  "both",
                     "Luniseta",    "both",
                     "Mdomestica",  "both",
                     "Mflaveola",   "both",
                     "Mscalaris",   "both",
                     "Sbullata",    "both",
                     "Scynipsea",   "both",
                     "Sstercorar",  "both"
                    );
```

```
my %cyclorappha = (  "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Cochliomyi",  "both",
                     "Dradicum",    "both",
                     "Dmelanogas",  "both",
                     "Ebalteatus",  "both",
                     "Elarvarum",   "both",
                     "Gmorsitans",  "both",
                     "Luniseta",    "both",
                     "Mdomestica",  "both",
                     "Mflaveola",   "both",
                     "Mscalaris",   "both",
                     "Sbullata",    "both",
                     "Scynipsea",   "both",
                     "Sstercorar",  "both"
                    );

my %platypez = (  "Luniseta",    "both",
                     "Mscalaris",   "both",
                    );

my %syrphschizo = (  "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Cochliomyi",  "both",
                     "Dradicum",    "both",
                     "Dmelanogas",  "both",
                     "Ebalteatus",  "both",
                     "Elarvarum",   "both",
                     "Gmorsitans",  "both",
                     "Mdomestica",  "both",
                     "Mflaveola",   "both",
                     "Sbullata",    "both",
                     "Scynipsea",   "both",
                     "Sstercorar",  "both"
                    );
```

```
my %schizophora = (  "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Cochliomyi",  "both",
                     "Dradicum",    "both",
                     "Dmelanogas",  "both",
                     "Elarvarum",   "both",
                     "Gmorsitans",  "both",
                     "Mdomestica",  "both",
                     "Mflaveola",   "both",
                     "Sbullata",    "both",
                     "Scynipsea",   "both",
                     "Sstercorar",  "both"
                    );

my %calyptratae = (  "Cochliomyi",  "both",
                     "Dradicum",    "both",
                     "Elarvarum",   "both",
                     "Gmorsitans",  "both",
                     "Mdomestica",  "both",
                     "Sbullata",    "both",
                     "Sstercorar",  "both"
                    );

#Note incompatable with schiz1
my %acalyptratae = ( "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Dmelanogas",  "both",
                     "Mflaveola",   "both",
                     "Scynipsea",   "both"
                    );
#Note incompatable with acalyptrate
my %schiz1 =      (  "Cochliomyi",  "both",
                     "Dmelanogas",  "both",
                     "Dradicum",    "both",
                     "Elarvarum",   "both",
```

```
                    "Gmorsitans",  "both",
                    "Mdomestica",  "both",
                    "Sbullata",    "both",
                    "Sstercorar",  "both"
                   );

#Note incompatable with acalyptrate
my %schiz2   =     ( "Ccapitata",   "both",
                     "Cdalmanni",   "both",
                     "Mflaveola",   "both",
                     "Scynipsea",   "both"
                    );

my %sepcer   =     ( "Ccapitata",   "both",
                     "Scynipsea",   "both"
                    );

my %mincyrt   =    ( "Cdalmanni",   "both",
                     "Mflaveola",   "both"
                    );

#Note: incompatible with Oest+Musc1, Oest+Musc2
my %muscoidea =   (  "Dradicum",    "both",
                     "Mdomestica",  "both",
                     "Sstercorar",  "both"
                    );

my %oestmusc1   = (  "Cochliomyi",  "both",
                     "Dradicum",    "both",
                     "Elarvarum",   "both",
                     "Sbullata",    "both",
                     "Sstercorar",  "both"
                    );


my %oestmusc2 =    (  "Cochliomyi",  "both",
```

```
                    "Dradicum",    "both",
                    "Elarvarum",   "both",
                    "Mdomestica",  "both",
                    "Sbullata",    "both",
                    "Sstercorar",  "both"
                   );

my %deliascat =   (  "Dradicum",    "both",
                    "Sstercorar",  "both"
                   );

my %oestroidea = (   "Cochliomyi",  "both",
                    "Elarvarum",   "both",
                    "Sbullata",    "both"
                   );

my %sarccoch = (     "Cochliomyi",  "both",
                    "Sbullata",    "both"
                   );

my %clade_hash = (   "Brachycera",              \%brachycera,
                    "Muscomorpha",             \%muscomorpha,
                    "Ortho",                   \%ortho,
                    "Ortho1",                  \%ortho1,
                    "Asiloidea",               \%asiloidea,
                    "Eremoneura",                     \%eremoneura,
                    "Cyclorappha",             \%cyclorappha,
                    "Platypezoidea",           \%platypez,
                    "Syrph+Schiz",                    \%syrphschizo,
                    "Schizophora",             \%schizophora,
                    "Calyptratae",             \%calyptratae,
                    "Acalyptratae",            \%acalyptratae,
                    "Schiz1",                  \%schiz1,
                    "Schiz2",                  \%schiz2,
                    "Sep+Cer",                 \%sepcer,
                    "Min+Cyrt",                \%mincyrt,
```

```
                        "muscoidea",                \%muscoidea,
                        "Oest+Musc2",               \%oestmusc2,
                        "Oest+Musc1",               \%oestmusc1,
                        "Delia+Scat",                       \%deliascat,
                        "Oestroidea",               \%oestroidea,
                        "Sarc+Coch",                \%sarccoch
                    );

   return \%clade_hash;
}

sub parse_date_file {
        my $date_file = shift;
        my $dates_hash;

        my %date_lookup = (  "22",          "Brachycera",
                             "23",          "Ortho",
                             "24",          "Asiloidea",
                             "25",          "Eremoneura",
                             "26",          "Cyclorappha",
                             "27",          "Platypezoidea",
                             "28",          "Syrph+Schiz",
                             "29",      "Schizophora",
                             "30",          "Sep+Cer",
                             "31",          "Schiz2",
                             "32",          "Schiz1",
                             "33",          "Calyptratae",
                             "34",          "Oest+Musc2",
                             "35",          "Oest+Musc1",
                             "36",          "Oestroidea",
                             "37",          "Sarc+Coch",
                             "38",          "Delia+Scat",
                             "39",          "Min+Cyrt"


                    );
```

```
        open (my $date_fh, "<", $date_file) or die $!;
        while (my $line = <$date_fh>) {
                chomp $line;
                unless ($line =~ m/\d/g) {
                        next;
                }
                my @ages = split (/\t/, $line);
                my $node = shift @ages;
                unless (exists $date_lookup{$node} ) {
                        next;
                }
                #get rid of min and max values
                foreach my $age (@ages) {
                        $age =~ s/\(.+\)//g;
                        $age =~ s/\s*//g;

                        #round value to nearest int
                        $age =~ m/(\d+\.*\d*)/;
                        $age = $1;
                        $age = int($age + 0.5);
                }
                $dates_hash -> {$date_lookup{$node}} = \@ages;
        }
        return $dates_hash;
}




sub test_node {
   my $tree = shift;
   my $clade_ref = shift;
   my $type = shift;
   $clade_ref =~ s/-bibio//gi;
```

```
my %trimmed_clade_hash = %$clade_ref;

foreach my $key (keys %$clade_ref) {
   if (($clade_ref -> {$key} eq "both") || ($clade_ref -> {$key} eq $type) ) {
      next;
   }
   else {
      delete $trimmed_clade_hash {$key};
      print "deleted $key\n";
   }
}

my @internals = @{$tree -> get_internals};


NODE:foreach my $node (@internals) {
   my @terminals = @{$node -> get_terminals};
   if (@terminals == 28) {
      my $number = @terminals;
      print "testing $node\tright number of taxa\t$number\n";
      foreach my $taxa (@terminals) {
         my $name;
         if ($taxa -> get_name) {
            if ($taxa -> get_name eq "Chominivor") {
               my $parent = $taxa -> get_parent;
               my $parent_name = $parent -> get_name;
               print "\t$parent_name is the parent\n";
            }
            $name = $taxa -> get_name;
            if ( ($name =~ m/(\d+)/i) && ($name < 100) ) {
               my @children = @{$taxa -> get_children};
               my $number_of_children = @children;
               print "\t$name has $number_of_children children\n";
               my $parent = $taxa -> get_parent;
               my $parent_name = $parent -> get_name;
```

```
                print "\t$parent_name is the parent\n";
             }
          }
          else {
             $name = "unknown";
          }
          print "Taxa\t$name\n";
          if (!exists $trimmed_clade_hash{$taxa -> get_name} ) {
             next NODE;
          }
       }
       return $node;
    }
    else {
       my $number = @terminals;
       next NODE;
    }
 }
}

#Added $clade name as parameter for divergence time estimate version
sub process_ancestor_node {

   my $node = shift;
   my $clade = shift;

   my $support = $node -> get_score;
   my $branch_length = $node -> get_branch_length;
   my $max_length = $node -> calc_max_path_to_tips;
   my $min_length = $node -> calc_min_path_to_tips;
   my $avg_length = calc_average_length ($node);

   #Added for fixed divergence time info
   $avg_length = $dates_hash -> {$clade} -> [$date_index];
```

```
   print $out_fh "\t$support\t$branch_length\t$avg_length\t$min_length\t$max_length";

}

sub calc_average_length {
   my $node = shift;
   my $sum = 0;

   my $num_terms = @{$node -> get_terminals};
   my @children = @{$node -> get_children};
   foreach my $child (@children) {
      descend_node ($child, "0", \$sum);
   }
   my $avg = $sum /= $num_terms;

   return $avg;
}
sub descend_node {
   my $node = shift;
   my $parental_length = shift;
   my $sum_ref = shift;
   my $branch_length = $node -> get_branch_length;
   $parental_length += $branch_length;

   if ($node -> is_terminal ){
      my $name = $node -> get_name;
      $$sum_ref += $parental_length;

   }
   else {
      my @children = @{$node -> get_children};
      foreach my $child (@children) {
         descend_node ($child, $parental_length, $sum_ref);
      }
   }
}
```

```
sub parse_file_name {
   my $file_name = shift;
   $file_name =~ s/(\..*)//gi;

   my @split = split (/-/, $file_name);
   my @returns = ($split[0],$split[4],$split[5],$split[6]);
   if ($returns[1] =~ m/aa/) {
      $returns[3] = $returns[2];
      $returns[2] = "NA";
   }

   return \@returns;
}
```

APPENDIX B “REPEAT_COUNT_6.PL”

```
#Program to identify tandem repeats in DNA sequences
#Identifies largest motifs first and determines if they can
#be decomposed into smaller repeats and then
#continues on to smaller motifs
#Script also calculates composition statistics in order
#to test significance of repeats (statistics not calculated
#within body of script.
#Script will function on DNA or mmino acid data

use strict;
use warnings;
use Data::Dumper;

# Maximum and minimum size of tandem motifs to detect
my $max_motif_size = 20;
my $min_motif_size = 1;

my $max_scattered_motif_size = 20;
my $min_scattered_motif_size = 2;

our $threshold = .8;



#Make script portable to dna
our $isdna = 1;
our @alphabet;
our $filler = "!";

if ($isdna == 1) {
        @alphabet = qw (A C T G);
}
else {
        @alphabet = qw (A C D E F G H I K L M N P Q R S T V W Y);
}
```

```
# #make lookup table to mask unacceptable characters
# our %accept;
# foreach my $character (@alphabet) {
#       $accept{$character} = 1;
# }



my @files = <*.fas>;

# TODO: Delete later
unlink "test.txt";

#contains all observed motifs for detection of scattered repeat motifs
#my $motifs;

foreach my $file (@files) {

        my %sequences;
        open (my $in_fh, "<", $file) or die $!;
        my $species;

        while (my $line = <$in_fh>) {
                chomp $line;
                unless ($line =~ m/\S/) {
                        next;
                }

                if ($line =~ m/^>/) {
                        $species = $line;
                        $species =~ s/^>//;
                }
                else {
                        my $sequence = uc $line;
```

```
                        $sequence =~ s/\s//;
                        #Just get rid of all non alphabet characters and replace with a
filler
                        $sequence =~ s/[^@alphabet]/$filler/g;



                        if (exists $sequences{$species}) {
                                $sequences{$species} .= $sequence;
                        }
                        else {

                                $sequences{$species} = $sequence;
                        }
                }
        }
        close $in_fh;

        my $outroot = $file;
        $outroot =~ s/\.fas//;

        Composition (\%sequences, $outroot);

        TandemCount ($max_motif_size, $min_motif_size, \%sequences, $outroot);

        WordCount   ($max_scattered_motif_size,   $min_scattered_motif_size,   \%sequences,
$outroot);
}
sub WordCount {
        my $max_size = shift;
        my $min_size = shift;
        my $sequences = shift;
        my $outroot = shift;
        #my $species_list;
        my $repeats;
        my $species_list_hash;
```

```
	#Count all remaining words in data set
	for (my $motif_length = $max_size; $motif_length >= $min_size; $motif_length--) {
		$|++;
		print "Identifying words of length $motif_length\n";
		foreach my $species (keys %$sequences) {
			$species_list_hash -> {$species} = 1;
			my $orig_sequence = uc $sequences -> {$species};
			for (my $i = 0; $i < $motif_length; $i++) {
				my $position = $i;
				my $sequence = substr ($orig_sequence, $i);
				my @working_sequence = $sequence =~ m/.{$motif_length}/g;
				foreach my $snippet (@working_sequence) {
					my @snippet = split (//, $snippet);
#					foreach my $char (@snippet) {
#						unless (exists $accept{$char} ) {
#							next SNIPPET;
#						}
#					}
					if ($snippet =~ m/\!/) {
						#print "skipping $snippet\n";
						next;
					}
					if (exists $repeats -> {$snippet} -> {$species}) {
						$repeats -> {$snippet} -> {$species} ++;
						$repeats -> {$snippet} -> {"total"} ++;

					}
					else {
						$repeats -> {$snippet} -> {$species} = 1;
						if (exists $repeats -> {$snippet} ->
{"total"}) {
							$repeats -> {$snippet} -> {"total"}
++;
						}
						else {
```

```
                                                $repeats -> {$snippet} -> {"total"} =
1;
                                                }
                                                $repeats  ->  {$snippet}  ->  {"length"}  =
$motif_length;
                                        }
                                }
                        }
                }
        }
        my @species_list = keys %$species_list_hash;
        @species_list = sort @species_list;
        #organize and print data
        PrintWords ($repeats, \@species_list, $outroot);

}
#Organizes and prints found word data
sub PrintWords {
        my $repeats = shift;
        my $species_list = shift;
        my $outroot = shift;

        #sort snippets by size and then by sequence
        my @snippets = keys (%$repeats);
        @snippets = sort {
                if ($repeats -> {$a} -> {"length"} > $repeats -> {$b} -> {"length"}) {
                        return -1;
                }
                elsif ($repeats -> {$a} -> {"length"} < $repeats -> {$b} -> {"length"}) {
                        return 1;
                }
                else {
                        return $a cmp $b;
                }
        } @snippets;
```

```
#       foreach (@snippets) {
#               print "$_\n";
#       }

        #Output results
        my $outfile = $outroot . ".wordcount.xls";
        open (my $out_fh, ">", $outfile) or die $!;
        #file header
        print $out_fh "motif\ttotal\tlength\t";
        foreach (@$species_list) {
                print $out_fh "$_\t";
        }
        print $out_fh "\n";

        #data
        foreach my $snippet (@snippets) {
                my $total = $repeats -> {$snippet} -> {"total"};
                my $length = $repeats -> {$snippet} -> {"length"};
                print $out_fh "$snippet\t$total\t$length\t";
                foreach my $species (@$species_list) {
                        if (exists $repeats -> {$snippet} -> {$species}) {
                                print $out_fh $repeats -> {$snippet} -> {$species} ."\t";
                        }
                        else {
                                print $out_fh "0\t";
                        }
                }
                print $out_fh "\n";
        }

}

sub TandemCount {
        my $max_motif_size = shift;
        my $min_motif_size = shift;
        my $sequences = shift;
```

```
        my $outroot = shift;
        my $out_fh;

        foreach my $species   (keys %$sequences) {
                print "Identifying tandem repeats in $species .. length:";
                my $out_file = $outroot . "-" . $species . ".xls";
                open ($out_fh, ">", $out_file) or die $!;
                print $out_fh "motif\tstart\tend\tlength\tperiod\trepetition\tsequence\n";
                for (my $motif_length = $max_motif_size; $motif_length >= $min_motif_size;
$motif_length--) {
                        $|++;
                        print " $motif_length";
                        IdentifyTandems ($motif_length, $species, $sequences, $out_fh);
                }
                print "\n";
        }
        close $out_fh;

}

sub IdentifyTandems {
        my $motif_length = shift;
        my $species = shift;
        my $sequences_ref = shift;
        my @char_array = split (//, $sequences_ref -> {$species});
        my $out_fh = shift;

        my $tandems;

        #For every motif
        MOTIF:for (my $i = 0; ($i < (@char_array - $motif_length)); $i++) {
                @char_array = split (//, $sequences_ref -> {$species});

                my $end = $i + $motif_length - 1;  # -1 in expression because dealing with
array indices
                my @motif = @char_array[$i .. $end];
```

```
		#if the motif matches my filler character "!", bail out and hit the next
motif
		foreach my $char (@motif) {
			if ($char eq "!") {
				next MOTIF;
			}
		}


		#Logic: Don't scan earlier in the sequence than we are currently at because
we have already done it.
		#		Grab next chunk of $motif_size because we are only
interested in tandem repeats, so the next chunk
		#		has to be an identical match to be at all interesting
		#		Use a sentinel to terminate the while loop when all repeats
found

		my $sentinel = 1;
		my $start = $i;
		while ($sentinel) {
			if ($end + $motif_length +1 >= @char_array) {
				$sentinel = 0;
			}
			my @next_slice = @char_array[$end + 1 .. $end + $motif_length];
			if (CompareArrays (\@motif, \@next_slice)) {
				$end += $motif_length;
				$i = $end;
			}
			else {
				$sentinel = 0;
			}
		}
		if ($start == $i) {
			next;
		}
		else {
```

```
                    ExtendMatch (\$start, \$end, \@motif, \@char_array);
                    $i = $end;
                    my @slice = @char_array[$start .. $end];
                    my $motif_ref = InternalSearch(\@motif, \@slice);
                    @motif = @$motif_ref;
                    #$motifs -> {join ('', @motif)} = 1;
                    print $out_fh join ('' , @motif);
                    print $out_fh "\t";
                    print $out_fh $start + 1;
                    print $out_fh "\t";
                    print $out_fh $end + 1;
                    print $out_fh "\t";
                    print $out_fh $end - $start +1;
                    print $out_fh "\t";
                    my $length = @motif;
                    print $out_fh "$length\t";
                    print $out_fh @slice/$length;
                    print $out_fh "\t";
                    print $out_fh PrettyPattern (\@motif, \@slice);
                    print $out_fh "\n";
                    ReplaceMatch($start, $end, \@char_array);
                    $sequences_ref -> {$species} = (join ('', @char_array));
#                   print "\n";
#                   print $sequences_ref -> {$species};
#                   print "\n";
#                   print $test_fh "@motif tandem $start .. $end\n";
               }


       }
}

# Very similar to "IdentifyTandems", but designed to return a new smaller motif size
```

```
# and test multiple motif sizes.  Should have built it all into IdentifyTandems, but whatever....
# ASSUMPTION:  Extension in either direction as part of a larger motif will capture all extensions
#                           required for smaller motifs as well.   Odd cases that involve a large repeat overlapping
#                           two small repeats on either side won't be caught, but large repeats take priority

sub InternalSearch {
        my $incoming_motif = shift;
        my $char_ref = shift;
        my @char_array = @$char_ref;

        # If we ever find a smaller motif that fills the whole character array we got, we update $return.
        # The smallest motif will be returned (we cound down from large to small.  If no smaller motif is found,
        # zero is returned
        my $return = $incoming_motif;

        #For each motif size...
        for (my $motif_length = @$incoming_motif; $motif_length >= $min_motif_size; $motif_length--) {
                #For every motif...
                for (my $i = 0; ($i < (@char_array - $motif_length)); $i++) {
                        my $end = $i + $motif_length - 1;   # -1 in expression because dealing with array indices
                        my @motif = @char_array[$i .. $end];

                        #Logic: Don't scan earlier in the sequence than we are currently at because we have already done it.
                        #               Grab next chunk of $motif_size because we are only interested in tandem repeats, so the next chunk
                        #               has to be an identical match to be at all interesting
```

```
                        #               Use a sentinel to terminate the while loop when all
repeats found

                        my $sentinel = 1;
                        my $start = $i;
                        while ($sentinel) {
                                if ($end + $motif_length +1 >= @char_array) {
                                        $sentinel = 0;
                                }
                                my   @next_slice   =   @char_array[$end   +   1   ..   $end   +
$motif_length];
                                if (CompareArrays (\@motif, \@next_slice)) {
                                        $end += $motif_length;
                                        $i = $end;
                                }
                                else {
                                        $sentinel = 0;
                                }
                        }
                        if ($start == $i) {
                                next;
                        }
                        else {
                                ExtendMatch (\$start, \$end, \@motif, \@char_array);
                                $i = $end;
                                if ($start == 0 && $end == (@char_array -1)) {
                                        $return = \@motif;
                                }

                        }
                }
        }
        return $return;
}
```

```
sub ExtendMatch {
	#Note!  These are ALL references, even the scalars!
	my ($start, $end, $motif, $char_array) = @_;

	#extend front
	my @reverse_motif = reverse @$motif;
	foreach my $char (@reverse_motif) {
		if ($$start == 0) {
			last;
		}
		if (($char_array -> [$$start - 1]) eq $char) {
			$$start--;
		}
		else {
			last;
		}
	}
	#extend rear
	foreach my $char (@$motif) {
		if ($$end == (@$char_array - 1)) {
			last;
		}
		if (($char_array -> [$$end + 1]) eq $char) {
			$$end++;
		}
		else {
			last;
		}
	}


}

# Subroutine replaces the specified sequence with an arbitrary filler character to
prevent future matches
sub ReplaceMatch {
```

```
        my $start = shift;
        my $end = shift;
        my $char_array = shift;       #Array ref
        my $sanity_check = @$char_array;

        #my $filler = "!";

        my @replace_array;
        for (my $i = $start; $i <= $end; $i++) {
                push (@replace_array, $filler);
        }

        my $length = $end-$start + 1;

        splice (@$char_array, $start, $length, @replace_array);

}



sub CompareArrays {
        my ($first, $second) = @_;
        #my $threshold = shift;
        my $match = 0;
        no warnings;  # silence spurious -w undef complaints
        return 0 unless @$first == @$second;
        for (my $i = 0; $i < @$first; $i++) {
                if ($first -> [$i] eq $second -> [$i]) {
                        $match++;
                }
            #return 0 if $first->[$i] ne $second->[$i];
        }
        if ($match/@$first >= $threshold) {
                #print "match\n";
                return 1;
        }
```

```
        else {
                return 0;
        }
}

sub Composition {
        my $sequences = shift;
        my $outroot = shift;
        my $compositions;
        my @sorted_species = sort keys %$sequences;
        print "Calculating site composition\n";
        foreach my $species (@sorted_species) {
                $compositions -> {$species} = CalcComp ($sequences -> {$species});
        }
        PrintComp ($outroot, \@sorted_species, $compositions);
}



sub CalcComp {
        my $sequence = shift;
        #my @alphabet = qw (A C D E F G H I K L M N P Q R S T V W Y);
        my %composition;
        my @sites = split (//, $sequence);
        my $length = @sites;
        foreach my $site (@sites) {
#               unless (exists $accept{$site} ) {
#                       next;
#               }
                if ($site =~ m/\!/) {
                        $length--;
                        next;
                }
                if (exists $composition{$site}) {
                        $composition{$site} ++;
                }
                else {
```

```
                $composition {$site} = 1;
        }
    }
    #Fill in missing AA and calc percentages
    foreach my $letter (@alphabet) {
        if (exists $composition{$letter}) {
            $composition{$letter} /= $length;
        }
        else {
            $composition{$letter} = 0;
        }
    }

    return \%composition;
}

# Takes ($outoot, \@sorted_species, $compositions)
sub PrintComp {
    my $outroot = shift;
    my $sorted_species = shift;
    my $compositions = shift;

    my $comp_outfile = $outroot . ".comp.xls";
    open (my $comp_fh, ">", $comp_outfile) or die $!;
    print $comp_fh "Site\t";
    foreach my $species (@$sorted_species) {
        print $comp_fh "$species\t";
    }
    print $comp_fh "\n";
    #my @alphabet = qw (A C D E F G H I K L M N P Q R S T V W Y);
    foreach my $site (@alphabet) {
        print $comp_fh "$site\t";
        foreach my $species (@$sorted_species) {
            print $comp_fh $compositions -> {$species} -> {$site};
            print $comp_fh "\t";
        }
```

```
                print $comp_fh "\n";
        }
        close $comp_fh;
}

sub PrettyPattern {
        my $motif = join ('', @{$_[0]});
        my $slice = join ('', @{$_[1]});
        my $length = @{$_[0]};

        #$motif = "APA";
        #$slice = "AAPAAPAPAPA";

        #$slice =~ s/$motif/ $motif /g;
        $slice =~ s/(.{$length})/$& /g;
        $slice =~ s/  / /g;
        $slice =~ s/(^ | $)//g;
        chomp $slice;
        return $slice;
}
```

REFERENCES

Aguinaldo, AMA, JM Turbeville, LS Linford, MC Rivera, JR Garey, RA Raff, JA Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-493.

Altekar, G, S Dwarkadas, JP Huelsenbeck, F Ronquist. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407-415.

Aris-Brosou, S. 2007. Dating Phylogenies with Hybrid Local Molecular Clocks. Plos One 2.

Baker, RH, GS Wilkinson, R DeSalle. 2001. Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Systematic Biology 50:87-105.

Birky, CW. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annual Review of Genetics 35:125-148.

Blagoderov, V, DA Grimaldi, NC Fraser. 2007. How time flies for flies: diverse Diptera from the Triassic of Virginia and early radiation of the order. American Museum Novitates:1-39.

Boore, J. 1999. Animal mitochondrial genomes. Nucl. Acids Res. 27:1767-1780.

Bourlat, SJ, T Juliusdottir, CJ Lowe, et al. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85-88.

Bourlat, SJ, C Nielsen, AD Economou, MJ Telford. 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Molecular Phylogenetics and Evolution 49:23-31.

Bowmaker, M, MY Yang, T Yasukawa, A Reyes, HT Jacobs, JA Huberman, IJ Holt. 2003. Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone. J. Biol. Chem. 278:50961-50969.

Brower, AVZ, R Desalle. 1994. Practical and theoretical considerations for choice of a DNA-sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. Annals of the Entomological Society of America 87:702-716.

Brown, WM, M George, AC Wilson. 1979. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 76:1967-1971.

Burger, G, MW Gray, B Franz Lang. 2003. Mitochondrial genomes: anything goes. Trends in Genetics 19:709-716.

Burleigh, JG, K Hilu, D Soltis. 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. BMC Evolutionary Biology 9:61.

Cameron, SL, SC Barker, MF Whiting. 2006. Mitochondrial genomics and the new insect order Mantophasmatodea. Molecular Phylogenetics and Evolution 38:274-279.

Cameron, SL, CL Lambkin, SC Barker, MF Whiting. 2007. A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. Systematic Entomology 32:40-59.

Caravas, J, M Friedrich. 2010. Of mites and millipedes: Recent progress in resolving the base of the arthropod tree. BioEssays 32:488-495.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Castresana, J, G Feldmaier-Fuchs, S Yokobori, N Satoh, S Paabo. 1998. The mitochondrial genome of the hemichordate Balanoglossus carnosus and the evolution of deuterostome mitochondria. Genetics 150:1115-1123.

Caterino, MS, RD Reed, MM Kuo, FAH Sperling. 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera : papilionidae). Systematic Biology 50:106-127.

Chvála, M. 1983. The Empidoidea (Diptera) of Fennoscandia and Denmark. II. General part. The families Hybotidae. Atelestidae, and Microphoridae. Fauna Entomol. Scand.:1-297.

Clayton, DA. 1982. Replication of animal mitochondrial DNA. Cell 28:693-705.

Conant, GC, PO Lewis. 2001. Effects of Nucleotide Composition Bias on the Success of the Parsimony Criterion in Phylogenetic Inference. Molecular Biology and Evolution 18:1024-1033.

Cumming, JM, Sinclair, B.J., Wood, D.M. 1995. Homology and phylogenetic implications of male genitalia in Diptera - Eremoneura. Entomol. Scand.:120-151.

Czelusniak, J, M Goodman, D Hewettemmett, ML Weiss, PJ Venta, RE Tashian. 1982. Phylogenetic origins and adaptive evolution of avian and mammalian hemoglobin genes. Nature 298:297-300.

Daly, M, LC Gusmao, AJ Reft, E Rodriguez. 2010. Phylogenetic signal in mitochondrial and nuclear markers in sea anemones (Cnidaria, Actiniaria). Integrative and Comparative Biology 50:371-388.

Douzery, EJP, F Delsuc, MJ Stanhope, D Huchon. 2003. Local molecular clocks in three nuclear genes: Divergence times for rodents and other mammals and incompatibility among fossil calibrations. Journal of Molecular Evolution 57:S201-S213.

Drummond, A, A Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214.

Drummond, AJ, SYW Ho, MJ Phillips, A Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88.

Drummond, AJ, MA Suchard. 2010. Bayesian random local clocks, or one rate to rule them all. Bmc Biology 8.

Dyer, NA, SP Lawton, S Ravel, KS Choi, MJ Lehane, AS Robinson, LM Okedi, MJR Hall, P Solano, MJ Donnelly. 2008. Molecular phylogenetics of tsetse flies (Diptera: Glossinidae) based on mitochondrial (COI, 16S, ND2) and nuclear ribosomal DNA sequences, with an emphasis on the palpalis group. Molecular Phylogenetics and Evolution 49:227-239.

Earl, D, K Bradnam, J St. John, et al. 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Research.

Enright, HU, WJ Miller, RP Hebbel. 1992. Nucleosomal histone protein protects DNA from iron-mediated damage. Nucleic Acids Research 20:3341-3346.

Evenhuis, N. 1994. Catalogue of the Fossil Flies of the World (Insecta: Diptera). Leiden: Backhuys.

Ewing, B, P Green. 1998. Base-calling of automated sequencer traces using Phred. II. error probabilities. Genome Research 8:186-194.

Ewing, B, L Hillier, MC Wendl, P Green. 1998. Base-calling of automated sequencer traces using Phred. I. accuracy assessment. Genome Research 8:175-185.

Fischer, M, M Steel. 2009. Sequence length bounds for resolving a deep phylogenetic divergence. Journal of Theoretical Biology 256:247-252.

Fisher, RA. 1922. On the mathematical foundation of theoretical statistics. Philosophical Transactions of the Royal Society of London:309-368.

Friedrich, M, D Tautz. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165-167.

Galewski, T, MK Tilak, S Sanchez, P Chevret, E Paradis, EJ Douzery. 2006. The evolutionary radiation of Arvicolinae rodents (voles and lemmings): relative contribution of nuclear and mitochondrial DNA phylogenies. BMC Evol Biol 6:80.

Gibson, A, V Gowri-Shankar, PG Higgs, M Rattray. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol 22:251-264.

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

Goodman, M. 1981a. Decoding the pattern of protein evolution. Progress in Biophysics & Molecular Biology 38:105-164.

Goodman, M. 1981b. Globin evolution was apparently very rapid in early vertebrates - A reasonable case against the rate-constancy hypothesis. Journal of Molecular Evolution 17:114-120.

Gredilla, R, VA Bohr, T Stevnsner. 2010. Mitochondrial DNA repair and association with aging - An update. Experimental Gerontology 45:478-488.

Griffiths, GCD. 1972. The Phylogenetic Classification of Diptera Cyclorrhapha, with Special Reference to the Structure of the Male Postabdomen. The Hague: Junk.

Griffiths, GCD. 1991. Book review of the Manual of Neartic Diptera, Vol 3. Quaest. Entomol.:117-130.

Griffiths, GCD. 1994. Relationships among the major subgroups of Brachycera (Diptera): a critical review. Can. Entomol.:861-880.

Grimaldi, D, MS Engel. 2005. Evolution of the Insects. New York: Cambridge University Press.

Hartmann, S, T Vision. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evolutionary Biology 8:95.

Hassanin, A. 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: Strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Molecular Phylogenetics and Evolution 38:100-116.

Hassanin, A, N Leger, J Deutsch. 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of metazoa, and consequences for phylogenetic inferences. Syst Biol 54:277-298.

Hennig, W. 1958. Die Familien der Diptera Schizophora und ihre phylogenetischen Verwandtschaftsbezeihungen. Beitr. Entomol.:505-688.

Hennig, W. 1971. Insektfossilien aus der unteren Kreide. III. Empidiformia ("Microphorinae") aus der untern Kreide und aus dem Baltischen Bernstein; ein Vertreter der Cyclorrhapha aus der unteren Kreide. Stuttg. Beitr. Naturkd:1-28.

Hennig, W. 1973. Diptera (Zweiflűgler). Handb. Zool. Berlin:1:200.

Hennig, W. 1981. Insect Phylogeny. New York: Wiley.

Huelsenbeck, JP, JJ Bull, CW Cunningham. 1996. Combining data in phylogenetic analysis. Trends in Ecology & Evolution 11:152-158.

Huelsenbeck, JP, F Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Hwang, UW, M Friedrich, D Tautz, CJ Park, W Kim. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413:154-157.

Jacobsen, F, NR Friedman, KE Omland. 2010. Congruence between nuclear and mitochondrial DNA: Combination of multiple nuclear introns resolves a well-supported phylogeny of New World orioles (Icterus). Molecular Phylogenetics and Evolution 56:419-427.

Jaillon, O, JM Aury, F Brunet, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946-957.

Jermiin, LS, SYW Ho, F Ababneh, J Robinson, AWD Larkum. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Systematic Biology 53:638-643.

Krivosheina, NP. 1989. Phyletic relations and evolution of the lower Brachycera (Diptera). Entomol. Obozr.:662-673.

Krivosheina, NP. 1991. Phylogeny of the lower Brachycera (Diptera). Acta Entomol. Bohemoslov:81-93.

Krivosheina, NP. 1998. Approaches to solutions of questions of classification of the Diptera. Entomol. Obozr.:378-390.

Krzemiński, WaK, E. 2003. Triassic Diptera: Descriptions, revisions, and phylogenetic relations. Acta Zoologica Cracoviensia:153-184.

Kumar, S, M Nei, J Dudley, K Tamura. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299-306.

Kutty, SN, T Pape, A Pont, BM Wiegmann, R Meier. 2008. The Muscoidea (Diptera: Calyptratae) are paraphyletic: Evidence from four mitochondrial and four nuclear genes. Molecular Phylogenetics and Evolution 49:639-652.

Labandeira, C. 1994. A compendium of fossil insect families. Milw. Public Mus. Contrib. Biol. Geol.:1-71.

Lepage, T, D Bryant, H Philippe, N Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669-2680.

Leys, R, SJB Cooper, MP Schwarz. 2002. Molecular phylogeny and historical biogeography of the large carpenter bees, genus Xylocopa (Hymenoptera : Apidae). Biological Journal of the Linnean Society 77:249-266.

Lin, C-P, BN Danforth. 2004. How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. Molecular Phylogenetics and Evolution 30:686-702.

Ljungman, M, PC Hanawalt. 1992. Efficient protection against oxidative DNA damage in chromatin. Molecular Carcinogenesis 5:264-269.

Margoliash, E. 1963. Primary structure and evolution of cytochrome c. Proceedings of the National Academy of Sciences of the United States of America 50:672-&.

McAlpine, JF, Peterson, B.V., Shewell, G.E., Teskey, H.J., Vockeroth, J.R., et al. 1981. Manual of Neartic Diptera. Ottawa, Can: Res. Branch. Agric. Can.

Michelson, V. 1996. Neodiptera - new insights into adult morphology and higher level phylogeny of Diptera (Insecta). Zool. J. Linn. Soc.:71-102.

Mindell, DP, MD Sorenson, DE Dimcheff, M Hasegawa, JC Ast, T Yuri. 1999. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. Systematic Biology 48:138-152.

Moriyama, E, J Powell. 1997. Synonymous substitution rates in <i>Drosophila</i> : Mitochondrial versus nuclear genes. Journal of Molecular Evolution 45:378-391.

Nagatomi, A. 1996. An essay on phylogeny of the orthorraphous Brachycera (Diptera). Entomol. Mon. Mag:95-148.

Oliveira, DCSG, R Raychoudhury, DV Lavrov, JH Werren. 2008. Rapidly Evolving Mitochondrial Genome and Directional Selection in Mitochondrial Genes in the Parasitic Wasp Nasonia (Hymenoptera: Pteromalidae). Molecular Biology and Evolution 25:2167-2180.

Oosterbroek P., CG. 1995. Phylogeny of the nematocerous families of Diptera (Insecta). Zool. J. Linn. Soc.:267-311.

Pamilo, P, M Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5:568-583.

Pape, T, FC Thompson. 2010. Systema Dipterorum. Version 1.0 http://www.diptera.org accessed on 7/11/11.

Phillips, MJ. 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441:132-140.

Rambaut, A, A Drummond. 2007. Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Systematic Biology 51:754-760.

Reed, RD, FAH Sperling. 1999. Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus Papilio. Molecular Biology and Evolution 16:286-297.

Rokas, A, SB Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337-1344.

Ronquist, F, JP Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rota-Stabelli, O, E Kayal, D Gleeson, J Daub, JL Boore, MJ Telford, D Pisani, M Blaxter, DV Lavrov. 2010. Ecdysozoan Mitogenomics: Evidence for a Common Origin of the Legged Invertebrates, the Panarthropoda. Genome Biology and Evolution 2:425-440.

Rubinoff, D, BS Holland. 2005. Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. Systematic Biology 54:952 - 961.

Sarich, VM, AC Wilson. 1967a. Immunological time scale for hominid evolution. Science 158:1200-1203.

Sarich, VM, AC Wilson. 1967b. Rates of albumin evolution in primates. Proc Natl Acad Sci U S A 58:142-148.

Silberfeld, T, JW Leigh, H Verbruggen, C Cruaud, B de Reviers, F Rousseau. 2010. A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): Investigating the evolutionary nature of the "brown algal crown radiation". Molecular Phylogenetics and Evolution 56:659-674.

Simon, C, F Frati, A Beckenbach, B Crespi, H Liu, P Flook. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers. Annals of the Entomological Society of America 87:651-701.

Sinclair, B. 1992. A phylogenetic interpretation of the Brachycera (Diptera) based on the larval mandible and associated mouthpart structures. Syst. Entomol:233-252.

Sinclair, BJ, Cumming, J.M., Wood, D.M. 1994. Homology and phylogenetic implications of male genitalia in Diptera - lower Brachycera. Entomol. Scand.:407-432.

Singh, B, H Kurahashi, JD Wells. 2011. Molecular phylogeny of the blowfly genus Chrysomya. Medical and Veterinary Entomology 25:126-134.

Song, S, ZF Pursell, WC Copeland, MJ Longley, TA Kunkel, CK Mathews. 2005. DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. Proceedings of the National Academy of Sciences of the United States of America 102:4990-4995.

Springer, MS, RW DeBry, C Douady, HM Amrine, O Madsen, WW de Jong, MJ Stanhope. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol Biol Evol 18:132-143.

Stajich, JE, D Block, K Boulez, et al. 2002. The Bioperl Toolkit: Perl modules for the life sciences. Genome Research 12:1611-1618.

Stewart, JB, AT Beckenbach. 2003. Phylogenetic and genomic analysis of the complete mitochondrial DNA sequence of the spotted asparagus beetle Crioceris duodecimpunctata. Molecular Phylogenetics and Evolution 26:513-526.

Stoffolano, JG, Woodley, N.E., Borkent, A., Yin, L.R.S. 1988. Ultrastructural studies of the abdominal plaques of some Diptera. Ann. Entomol. Soc. Am.:503-510.

Strope, CL, K Abel, SD Scott, EN Moriyama. 2009. Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses: indel-Seq-Gen Version 2.0. Molecular Biology and Evolution 26:2581-2593.

Sullivan, J, P Joyce. 2005. Model selection in phylogenetics. Annual Review of Ecology Evolution and Systematics 36:445-466.

Svennblad, B. 2008. Consistent Estimation of Divergence Times in Phylogenetic Trees with Local Molecular Clocks. Systematic Biology 57:947-954.

Talavera, G, J Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.

Townsend, JP. 2007. Profiling phylogenetic informativeness. Systematic Biology 56:222-231.

Vos, R, J Caravas, K Hartmann, M Jensen, C Miller. 2011. BIO::Phylo-phyloinformatic analysis using perl. BMC Bioinformatics 12:63.

Wada, S. 1991. Morphologische Indizien für das unmittelbare Schwester-gruppenverhältnis der Schizophora mit den Syrphoidea (‘Aschiza’) in der phylogenetischen Systematik der Cyclorrhapha (Diptera: Brachycera). J. Nat. Hist.:1531-1570.

Waddell, PJ, Y Cao, J Hauf, M Hasegawa. 1999. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid invariant sites LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Systematic Biology 48:31-53.

Webster, BL, RR Copley, RA Jenner, JA Mackenzie-Dodds, SJ Bourlat, O Rota-Stabelli, DTJ Littlewood, MJ Telford. 2006. Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evolution & Development 8:502-510.

Wiegmann, BM, Mitter, C., Thompson, F.C. 1993. Evolutionary origin of the Cyclorrapha (Diptera): tests of alternative morphological hypotheses. Cladistics:41-81.

Wiegmann, BM, MD Trautwein, IS Winkler, et al. 2011. Episodic radiations in the fly tree of life. Proceedings of the National Academy of Sciences 108:5690-5695.

Wiegmann, BM, DK Yeates, JL Thorne, H Kishino. 2003. Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. Systematic Biology 52:745-756.

Wiens, JJ. 1998. Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy? Systematic Biology 47:625-640.

Willman, R. 1989. Evolution und phylogenetisches System der Mecoptera (Insecta, Holometabola). Abhandlungen der Senckenbergischen Naturforschenden Gesellschaft:1-153.

Wilson, AC, VM Sarich. 1969. A molecular time scale for human evolution. Proc Natl Acad Sci U S A 63:1088-1093.

Winkler, IS, CC Labandeira, T Wappler, P Wilf. 2010. Distinguishing Agromyzidae (Diptera) leaf mines in the fossil record: new taxa from the paleogene of North America and Germany and their evolutionary implications. Journal of Paleontology 84:935-954.

Wood, DM. 1991. Homology and phylogenetic implications of male genitalia in Dipter. The ground plan. In: OI Weismaan L, editor. Proc. 2nd Int Congr. Dipterology. Bratislava, Czechoslovakia: The Hague: Academic. p. 255-284.

Wood, VR GwilliamMA Rajandream, et al. 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415:871-880.

Woodley, N. 1989a. Phylogeny and classification of the "Orthorraphous" Brachycera. In: WD McAlpine JF, editor. Manual of the Neartic Diptera. Ottawa, Can.: Res. Branch Agric. Can. p. 1371-1395.

Woodley, N. 1989b. Phylogeny and classification of the "Orthorrhaphous" Brachycera. In: J Mc Alpine, D Woods, editors. Manual of the Neartic Diptera. Ottawa, Can: Res. Branch Agric. Can. p. 1371-1395.

Wray, GA, JS Levinton, LH Shapiro. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. Science 274:568-573.

Wu, J, E Susko, AJ Roger. 2008. An independent heterotachy model and its implications for phylogeny and divergence time estimation. Molecular Phylogenetics and Evolution 46:801-806.

Yang, Z, B Rannala. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23:212-226.

Yang, Z, AD Yoder. 2003. Comparison of Likelihood and Bayesian Methods for Estimating Divergence Times Using Multiple Gene Loci and Calibration Points, with Application to a Radiation of Cute-Looking Mouse Lemur Species. Systematic Biology 52:705-716.

Yang, ZH. 1998. On the best evolutionary rate for phylogenetic analysis. Systematic Biology 47:125-133.

Yeates, DK, BM Wiegmann. 1999. Congruence and controversy: Toward a higher-level phylogeny of diptera. Annual Review of Entomology 44:397-428.

Yoder, AD, ZH Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Molecular Biology and Evolution 17:1081-1090.

Zatwarnicki, T. 1996. A new reconstruction of the origin of eremoneuran hypopygium and its implications for classification (Insecta: Diptera). Genus:103-175.

Zhong, J, J Zhang, E Mukwaya, Y Wang. 2009. Revaluation of deuterostome phylogeny and evolutionary relationships among chordate subphyla using mitogenome data. Journal of Genetics and Genomics 36:151-160.

Zink, RM, GF Barrowclough. 2008. Mitochondrial DNA under siege in avian phylogeography. Molecular Ecology 17:2107-2121.

Zuckerkandl, E, L Pauling. 1965. Evolutionary divergence and convergence in proteins. In: V Bryson, HJ Vogel, editors. Evolving Genes and Proteins. New York: Academic Press. p. 97-166.

Zuckerkandl, E, LB Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. In: M Kasha, B Pullman, editors. Horizons in Biochemistry: Academic Press. p. 189-225.

ABSTRACT

PHYLOGENETIC UTILITY OF MITOCHONDRIAL AND NUCLEAR GENES: A CASE STUDY IN THE DIPTERA (TRUE FLIES)

by

JASON CARAVAS

May 2012

Advisor: Dr. Markus Friedrich

Major: Biological Sciences

Degree: Doctor of Philosophy

The value of mitochondrial versus nuclear gene sequence data in phylogenetic analysis has received much attention without yielding definitive conclusions. Theoretical arguments and empirical data suggest a lower phylogenetic utility than equivalent nuclear gene sequences, but there are also many examples of important progress made using mitochondrial sequences. We therefore undertook a systematic performance analysis of mitochondrial and nuclear sequence partitions taken from a representative sample of dipteran species. For phylogenetic tree reconstruction, mitochondrial genes performed generally inferior to nuclear genes. However, the mitochondrial genes resolved branches for which nuclear genes failed. Moreover, the combined use of mitochondrial and nuclear sequences produced superior results without artifacts for nodes where mitochondrial and nuclear gene data sets on their own generated conflicting topologies. These findings strongly advocate the inclusion of mitochondrial sequences even in deep phylogeny reconstruction. The comparison of tree support between our and previous analyses identified robustly supported high confidence clades in the Diptera but also a

number of problematic groupings in need of further analysis. For divergence time estimation, we show widespread convergence of clade age estimates from both mitochondrial and nuclear gene sources under a wide variety of data preparation and model paradigms. Our results indicate slightly superior performance of nuclear gene derived ages for nodes for several clades in the tree ranging in age from approximately 30 to 160 myo. We further find that third codon position inclusion negatively affects our ability to resolve ages under many circumstances. Increasing model complexity and granularity of data partitioning offered little benefit in terms of final results while increasing the computational complexity. Finally, we produce high confidence age estimates for cyclorrhaphan divergences in agreement with previous literature, and provide the first timeline for major divergences within the calyptrate flies.

AUTOBIOGRAPHICAL STATEMENT

Jason Caravas received a Bachelor of Science degree in Biological Science from Wayne State University in 1999. He enrolled in the Wayne State University PhD program in 2002 under the mentorship of Dr. Markus Friedrich. During his PhD studies, Jason participated in the Assembling the Tree of Life: Diptera project and investigated the strength of phylogenetic signal in mitochondrial and nuclear genes using the Diptera as a test data set. Other projects included investigation of gene duplication events among arthropod groups and analysis of high throughput sequence reads for the purpose of assembling a transcriptome of the cave beetle *Ptomophagus hirtus*. While working in Dr. Friedrich's lab, Jason also acquired a Certificate in Scientific Computing from Wayne State University. He developed programming skills, including grid computing and a contribution to the Bio::Phylo Perl module, during this period.

Jason Caravas received an NSF Interdisciplinary Graduate Research Traineeship (IGERT) fellowship, a Thomas C. Rumble competitive fellowship, and several Wayne State GRA Enhancement fellowships during his time in the PhD program at Wayne State. He presented at an international conference, the 6th International Congress of Dipterology 2006 (Fukuoka, Japan). He also participated in the Workshop on Molecular Evolution (Woods Hole Oceanographic Institute, MA), the Computational Phyloinformatics Workshop (National Evolutionary Synthesis Center, NC), and a Google Summer of Code project for the development of an XML based file format (NeXML) for storing phylogenetic information (National Evolutionary Synthesis Center, NC).