LARGE DATASETS AND TRICHOPTERA PHYLOGENETICS: DNA BARCODES, PARTITIONED PHYLOGENETIC MODELS, AND THE EVOLUTION OF PHRYGANEIDAE

By

PAUL BRYAN FRANDSEN

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Entomology

Written under the direction of

Karl M. Kjer

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER 2015

ABSTRACT OF THE DISSERTATION

Large datasets and Trichoptera phylogenetics: DNA barcodes, partitioned phylogenetic models, and the evolution of Phryganeidae

By PAUL BRYAN FRANDSEN

Dissertation Director:

Karl M. Kjer

Large datasets in phylogenetics—those with a large number of taxa, e.g. DNA barcode data sets, and those with a large amount of sequence data per taxon, e.g. data sets generated from high throughput sequencing—pose both exciting possibilities and interesting analytical problems. The analysis of both types of large datasets is explored in this dissertation. First, the use of DNA barcodes in phylogenetics is investigated via the generation of phylogenetic trees for known monophyletic clades. Barcodes are found to be useful in shallow scale phylogenetic analyses when given a well-supported scaffold on which to place them. One of the analytical challenges posed by large phylogenetic datasets is the selection of appropriate partitioned models of molecular evolution. The most commonly used model partitioning strategies can fail to characterize the true variation of the evolutionary process and this effect can be exacerbated when applied to large datasets. A new, scalable algorithm for the automatic selection

of partitioned models of molecular evolution is proposed with an eye toward reducing systematic error in phylogenomics. The new algorithm is tested on a range of empirical datasets and found to provide a better fit of the model to the data as measured by information theoretic metrics like AICc. Indeed, the algorithm is found to perform particularly well when applied to a phylogenomic dataset consisting of ultra-conserved elements (UCEs). Finally, the phylogeny of Phryganeidae is estimated using a large dataset generated using targeted enrichment and high throughput sequencing. Trees generated from different modeling strategies give incongruent, but strongly supported results. The differences between the trees are examined and a new hypothesis for the relationships among the genera within Phryganeidae is posited.

ACKNOWLEDGMENTS

This dissertation contains two chapters that were previously peer reviewed and are either published or in press. These include Chapter 1 (Frandsen et al. *in press*) and Chapter 2 (Frandsen et al. 2015). I wrote both papers with the help of several co-authors. I thank them for their contributions.

I would like to thank the sources that provided funding during my PhD: the Rutgers School of Environmental and Biological Sciences Excellence Fellowship, NSF DEB 0816865, the Rutgers Department of Entomology Thomas J. Headlee Fellowship, the National Evolutionary Synthesis Center and Google for the Phyloinformatics Google Summer of Code internship, and the German Academic Exchange Service (DAAD) for short term fellowship support.

There are many people who have influenced my life and helped enable me to complete my PhD. It would be impossible to give thanks to all. First, thank you to my adviser, Dr. Karl Kjer, for his advice, mentorship, and support during my PhD. The way I approach science has been deeply affected by our relationship and your guidance. Thank you to my committee members, Dr. Dina Fonseca, Dr. Ralph Holzenthal, and Dr. Mike May. Thank you to Dina for pushing back when I needed it, thanks to Ralph for the encouragement to be meticulous (even though I often failed in that regard), and thank you to Mike for encouraging and augmenting my interest in the fascinating world of organismal entomology. Thank you to the Rutgers Department of Entomology: the faculty, staff, and my

fellow graduate students. Thank you to Dr. Robert Lanfear, Dr. Brett Calcott, Dr. Christoph Mayer, and Dr. Bernhard Misof for the mentorship and guidance in bioinformatics training. Thank you to my undergraduate mentor, Dr. Riley Nelson, for instilling in me a sense of wonder that helped lead me into this path. Thank you to my many excellent siblings, Valerie, Mindy, Devn, Justin, Camie, and Jeffrey. Given that I am the 6th of 7 children, you are all just as responsible for who I am as anyone else. Thank you to my endlessly supportive mother, Donna. I am forever grateful for the fortune of being one of your sons. Thank you to my late father, Greg. You instilled in me an appreciation for living things and taught me to take advantages of my strengths and fight against my weaknesses. Lastly, thank you to my little family to whom this dissertation is dedicated: my wonderful and supportive wife, Christine—this journey would have been impossible without you, and my new son, Harvey Bryan—your precious smiles were all the encouragement that I needed to finish this thing.

References

Frandsen, Paul B., Brett Calcott, Christoph Mayer, and Robert Lanfear. 2015. "Automatic Selection of Partitioning Schemes for Phylogenetic Analyses Using Iterative K-Means Clustering of Site Rates." *BMC Evolutionary Biology* 15 (1): 13. doi:10.1186/s12862-015-0283-7.

Frandsen, Paul B., Oliver Flint, Xin Zhou, and Karl M. Kjer. *in press*. "A Proposal for Using Barcode Data to Fill out the Leaves on the Trichoptera Tree of Life." *Proceedings of the XIVth International Symposium on Trichoptera*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION TO DISSERTATION

Humans have long held the desire to classify and name things. This inclination is both experimental and practical. We are curious creatures and our world holds much to fear, consume, and learn about. The classification of the living things around us fulfills our innate desire to put things in order. It also fulfills a practical need in helping us to identify organisms that can be useful and to avoid those that are harmful. In the early 18th century, the process of naming living things was formalized with Linnaeus' publication of the 10th edition of *Systema Naturae* (Linnaeus 1758) and corresponding invention of the system of binomial nomenclature. These conventions have survived and, despite a decline in the hiring of taxonomists (Wheeler 2014), taxonomy is still a robust science, producing many new species descriptions every year. A natural extension of taxonomy is to contextualize named living things into the framework of groups, or classifications. Linnaeus did this by grouping similar creatures together by type into genera and other higher groupings such as families and orders. However, the nature of these hierarchical groupings were not fully understood until Charles Darwin posited the underlying relationships among all living things and their corresponding common ancestry, giving context to all of comparative biology (Darwin 1859). Following the widespread acceptance of evolution among scientists, the hierarchical structure gained meaning as a grand tree of life with common ancestors begetting the subsequent branches and leaves of the tree.

The German dipterist, William Hennig, formalized our phylogenetic system of classification by defining the cladistic convention of naming only monophyletic groups and establishing the evolutionary context behind this new discipline of phylogenetic systematics (Hennig 1966). Since then, the science of systematics and phylogenetics has undergone a revolution brought on by DNA sequencing technology. Many of the top 100 cited scientific papers of all time are molecular phylogenetics methods papers (Van Noorden, Maher, and Nuzzo 2014; Saitou and Nei 1987; Felsenstein 1985; Tamura et al. 2007; Posada and Crandall 2001; Ronquist and Huelsenbeck 2003). With the advent of high throughput DNA sequencing, phylogenetics has also become increasingly important for methods development in the nascent field of genomics. Phylogenetics has entered its adolescent phase, and it has debuted in a significant way.

Although the science of organizing and classifying living organisms has undergone major revolutions with incredibly powerful new tools available capable of producing more data than we could have previously imagined, we are still a long way from resolving the tree of life. Systematics, the science of classifying organisms; phylogenetics, the science of building a tree to put those classifications into an evolutionary context; and taxonomy, the science of naming things, must work hand in hand to reconstruct a comprehensive view of the evolutionary history of organisms on earth. Each field relies heavily on the others. After all, how might we organize the evolutionary tree of life without the discovery of the organisms we wish to classify?

This dissertation delves into the phylogenetics and evolutionary history of one order of insects, the caddisflies (Insecta: Trichoptera). Although taxonomy isn't the focus of this dissertation, there can be no evolutionary context for a set of organisms if those organisms have not yet been discovered. Therefore, none of this research could have been possible without the tireless efforts and expertise of a slew of Trichoptera taxonomists.

The generation of large amounts of molecular data is changing the way phylogenetics is conducted. Both DNA barcoding and the new era of high throughput sequencing represent areas where “big data” is being generated. In the case of DNA barcodes, a small amount of data is being generated for an enormous amount of individuals, while high throughput sequencing is generating large amounts of data for fewer individuals. Both cases offer exciting avenues to explore and set the theme for the following research.

I begin this dissertation with work related to the DNA barcoding effort (Chapter 1). This ambitious program has worked toward collecting and sequencing the DNA barcoding gene for as many living things as possible (Hebert et al. 2003). From a practical standpoint, no serious effort is put into the barcoding of a particular group without a dedicated expert in charge. From 2007-2010, Xin Zhou led the Trichoptera barcode of life initiative, handling samples from his own collecting in China and Canada, as well as receiving samples from the

collections of the University of Minnesota, the Smithsonian collection, and worldwide collaborators. I focus on this massive effort within Trichoptera and give recommendations on how these data can be used in phylogenetics when given an evolutionary scaffold. I then provide a real world example of how tree building with DNA barcodes can unveil interesting findings for museum collections. This anecdote is written in an informal style to convey the importance of stories and collaborations in scientific training and discovery.

During the course of my dissertation, I spent the majority of my effort toward understanding the mathematical models used to infer phylogenetic trees. In chapter 2, I outline a new bioinformatics algorithm that I developed to help researchers automatically select partitioned models of molecular evolution for their data. I show that, when evaluated using information theoretic metrics, automatically selected phenomenological models can contain more information than the more commonly used, mechanistic models that rely on simplified assumptions of the biological process. I also point out an inherent bias in previous studies that used a single starting tree to estimate site rates for partitioning—that such partitioning strategies can bias the final tree toward the starting tree.

Finally, I investigate the phylogenetic history of the caddisfly family Phryganeidae using recent technological advances in targeted enrichment and high throughput sequencing (Chapter 3). I also explore the use of partitioned

models and data exclusion with a large phylogenetic dataset. As the field crosses into the "big data" era of phylogenetics, we must spend time examining current phylogenetic methods and explore the reasons that we recover highly supported, conflicting answers using different methods. Every scientific enterprise experiences growing pains as it transitions into a data rich environment and these periods allow for the exploration of many important and potentially transformative questions. This exciting time is the state of the field of phylogenetics as I complete my PhD research.

References

Darwin, Charles. 1859. *On the Origin of Species*. 1st ed. London: John Murray.

Felsenstein, Joseph. 1985. “Confidence Limits on Phylogenies: An Approach Using the Bootstrap.” *Evolution* 39: 783–91. doi:10.2307/2408678.

Hebert, Paul D N, Alina Cywinska, Shelley L Ball, and Jeremy R deWaard. 2003. “Biological Identifications through DNA Barcodes.” *Proceedings. Biological Sciences / The Royal Society* 270 (1512): 313–21. doi:10.1098/rspb.2002.2218.

Hennig, Willi. 1966. *Phylogenetic Systematics*.

Linnaeus, Carl. 1758. *Systema Naturae per Regna Tira Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis.* 10th ed. Vol. 1, Animalia. Stockholmiae: Laurentii Salvii.

Posada, David, and Keith A. Crandall. 2001. “Selecting the Best-Fit Model of Nucleotide Substitution.” *Systematic Biology* 50 (4): 580–601. doi:10.1080/106351501180469.

Ronquist, Fredrik, and John P. Huelsenbeck. 2003. “MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models.” *Bioinformatics* 19 (12): 1572–74. doi:10.1093/bioinformatics/btg180.

Saitou, N., and M. Nei. 1987. “The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.” *Molecular Biology and Evolution* 4 (4): 406–25.

Tamura, Koichiro, Joel Dudley, Masatoshi Nei, and Sudhir Kumar. 2007. “MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0.” *Molecular Biology and Evolution* 24 (8): 1596–99. doi:10.1093/molbev/msm092.

Van Noorden, Richard, Brendan Maher, and Regina Nuzzo. 2014. “The Top 100 Papers.” *Nature* 514 (7524): 550–53. doi:10.1038/514550a.

Wheeler, Quentin. 2014. “Are Reports of the Death of Taxonomy an Exaggeration?” *New Phytologist* 201 (2): 370–71. doi:10.1111/nph.12612.

CHAPTER 1: USING DNA BARCODE DATA TO ADD LEAVES TO THE TRICOHPTERA TREE OF LIFE

Abstract

The Trichoptera barcode of life initiative has gathered barcodes for a large portion of Trichoptera species diversity. Although the primary use of these data is species identification, they can also be used to generate species-level phylogenetic hypotheses. In order to ameliorate the well-documented difficulties of resolving deep divergences with the COI barcode fragment, we used a method of defining well-supported nodes from other data sources and filling out the "leaves" within these defined nodes of the Trichoptera tree of life with trees generated from the barcode fragment. We demonstrate the potential of this approach with the generation of a tree for Xiphocentronidae + Psychomyiidae. Using this example, we present two suspicious clades that warranted a more careful analysis, and demonstrate that a simple analysis of barcodes can generate and help answer other, related questions. We find that *Zelandoptila* is supported as belonging to Ecnomidae rather than Psychomyiidae, and we placed an unidentified specimen from the Smithsonian National Museum of Natural History collection as sister to Beraeidae.

Introduction

A large, multidisciplinary effort has gone into DNA barcoding since the recommendation by Hebert *et al.* (2003a), to use a portion of the mitochondrial cytochrome oxidase subunit 1 (COI) gene as the standard "barcode" for

animals. The purpose of the barcode of life is to aid in species-level identification (Hebert *et al.* 2003b), which is important for identifying trichopteran larvae, and particularly useful for non-specialists. The barcode initiative does not advance at random, and one of the administrative requirements for the selection of target groups is the dedication of a specific taxonomic expert, and a community of researchers who are willing to help. Xin Zhou worked for four years at the Biodiversity Institute of Ontario in Guelph, Canada, leading the Trichoptera Barcode of Life initiative (http://trichopterabol.org/). Our community has dedicated countless hours of collecting, identifying, imaging, curating, and sequencing toward making Trichoptera among the best-characterized taxa on Earth.

Biodiversity distributions at every taxonomic level follow hollow curve distributions (Willis & Yule 1922), with a relatively few abundant taxa, such as Coleoptera, or Hydropsychidae, or *Hydropsyche*, and many that are rare, such as Embioptera, or Barbarochthonidae, or *Barbarachthon*. In other words, wherever you are in the world, you are more likely to be able to collect a hydropsychid than a barbarachthonid. The Barcode of Life Data systems (BOLD, http://www.boldsystems.org) contain over 44,000 specimen records for Trichoptera, including >4,800 species, representing all 48 families (accessed 17 Jan. 2013). With approximately 2/3^{rds} of all genera, and 1/3rd of all described species, most species, randomly encountered, are already represented in the database, because the abundant species have been collected. Although seemingly counter-intuitive, 1/3rd of the species represent the majority of species

we encounter. The ability to identify Trichoptera is particularly important for freshwater biomonitoring (Resh & Unzicker 1975). Many larvae have not been associated with their adult counterparts and species identification of larvae can be difficult or impossible. Enabled with barcodes, however, researchers can place larvae within a species and improve the resolution of biomonitoring data (Ruiter *et al.* 2013; Sweeney *et al.* 2011; Zhou *et al.* 2007). Barcodes can also confirm morphological hypotheses of species distinctiveness, as shown by Flint & Kjer (2011). As useful as barcode data can be, these sequences can only be associated to biological data if they are tied to a species name that has been provided by a taxonomic expert.

Although clearly useful for biomonitoring, barcodes have seldom been utilized for phylogenetics. One of its problems can be illustrated by the following comparison. Completely discreet characters are both ideal and rare for morphological phylogenetics. However, all characters in DNA are discreet, and there are only 4 of them. This means that with time, the same sites are likely to experience multiple changes, with reversals and parallel changes being indistinguishable from synapomorphies. If you wish to explore deep nodes in a molecular phylogeny, it is important to select a gene that has not experienced multiple superimposed substitutions. Different genes evolve at different rates, and an ideal barcode gene, with measurable variation even among populations within the same species is, by nature, a particularly poor gene for deep level phylogenetics. This was confirmed by Kjer *et al.* (2001; Figs. 3, 4, and 9), who

showed that pairwise sequence differences (the percent of nucleotides that differ between 2 taxa) are already at a maximum level within trichopteran suborders, and do not increase, even as you compare between orders. If change is occurring at some rate over time, the fact that the pairwise difference between a philopotamid and a hydropsychid is as great as the pairwise difference between a philopotamid and a mecopteran means that the same sites are changing over and over again, and the COI is not an ideal marker for estimating the relationships among trichopteran suborders. Therefore, some strategy is necessary for using barcode data to generate phylogenies.

There is a tradeoff between the amount of data available, and the number of taxa for a particular set of genes. Through the 1000 Insect Transcriptome Evolution initiative (1KITE, http://1kite.org/), we are developing transcriptomes for about 25 of the 48 trichopteran families. These data will be used to infer the relationships among suborders, with particular reference to the five "spicipalpian" families. At the next level, we have about 8000 nucleotides, from 5 genes (18S, 28S, EF1a, CAD, and COI) for 250 taxa representing all of the families, and most of the genera. Next, we have approximately 1000 nucleotides of 28S ribosomal RNA data for over 1200 individuals. Then we have the 658 nucleotides of barcode data from 40,000 individuals. Our strategy is to subdivide the order into well established clades, as close to the tips as possible, using the more conservative markers, and then use the barcode data to fill out the leaves of the tree, while leaving the trunk and basal branches to other, more appropriate markers. This strategy is based upon the observation that COI is

variable among recent divergences, and homoplastic as divergences go deeper. Here we present an example of how our strategy could work.

Materials and Methods

We sifted through all 40,000 trichopteran haplotypes in the barcode of life database (BOLD), and generated trees for well-supported, monophyletic nodes within the trichopteran phylogeny. In the process of constructing these trees and generating hypotheses about the relationships among trichopteran taxa, we can gain insights into ways to improve on the existing classification and solve difficult taxonomic problems. We highlight two such circumstances that occurred as we worked to generate hypotheses about the Trichoptera tree of life.

We decided to conduct a test of our strategy of filling out the leaves of the Trichoptera tree with two relatively small families, Xiphocentronidae and Psychomyiidae. These families have been proposed to be sister taxa by multiple studies, using both molecular and morphological data (Frania & Wiggins 1997; Holzenthal *et al.* 2007; Kjer *et al.* 2001, 2002; Kjer *et al*., this symposium volume). Sequences for Psychomyiidae and Xiphocentronidae were downloaded from the BOLD website. The datasets were filtered to include only sequences greater than 350 bp, lacking stop codons and those that were not flagged as contaminants or misidentifications as applied by the BOLD website. Identical haplotypes were combined so that each terminal on the tree represented a

unique haplotype. The resulting sequences were imported into Seaview (Gouy *et al.* 2010) and aligned using MUSCLE (Edgar 2004). Alignments were then inspected by eye.

A pseudoreplicate site-specific rate model was used to partition the data described by Kjer (Kjer *et al.* 2001; Kjer & Honeycutt 2007). Character partitions were constructed in PAUP* 4.0 (Swofford 2003). All analyses were run using RAxML 7.4 (Stamatakis *et al.* 2005). 1,000 rapid bootstraps were performed followed by a maximum likelihood tree search using the GTRGAMMA model. Trees were viewed and exported using FigTree (Rambaut 2012) then edited using Adobe Illustrator CS5 (2010). Suspect sequences were checked for contaminants against the GenBank database using nucleotide BLAST (Altschul *et al.* 1990). Other datasets were analyzed in the same fashion as the Psychomyiidae and Xiphocentronidae dataset.

Results and Discussion

Our phylogeny is shown in Fig. 1. However, this final result could not have been recovered in an automated analysis without careful data curation. When analyzing the initial tree output from the Psychomyiidae + Xiphocentronidae tree, the extreme branch lengths associated with the two “Xiphocentronidae sp.” taxa and the six *Zelandoptila* taxa were conspicuous (Fig. 2). Long branches are often indicative of a contaminant. Whenever such apomorphic sequences are observed, it is wise to search through GenBank for a close match with a

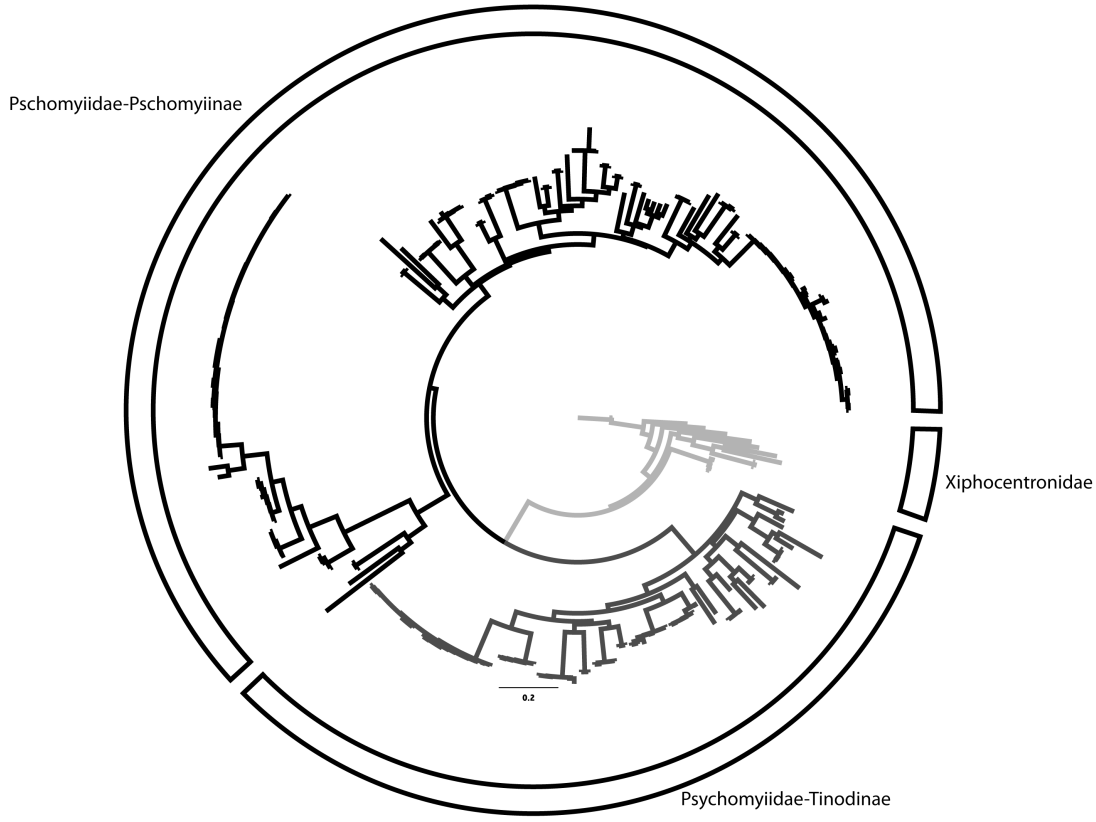

Figure 1. Phylogram from RAxML for the final analysis of the Psychomyiidae and Xiphocentronidae dataset.

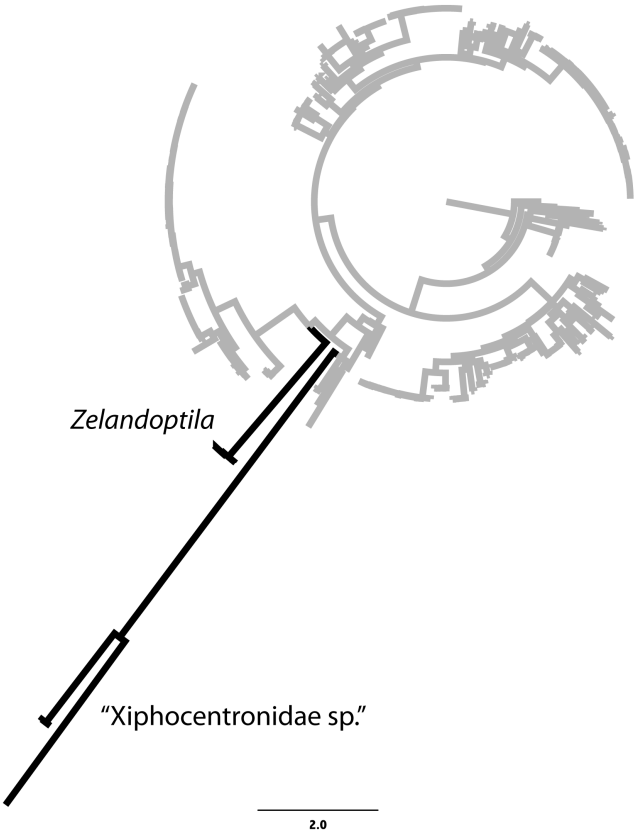

Figure 2. Phylogram from RAxML of the analysis of the output from BOLD for Xiphocentronidae and Psychomyiidae. Two unusually long branch lengths are present for "Xiphocentronidae sp." and *Zelandoptila*.

"BLAST" search. For example, if an unusual sequence in a Trichoptera study is an identical match to human DNA, then it is likely that you are looking at your own DNA. The BLAST search in this case returned several Trichoptera that were closely related to both taxa so the possibility of a non-trichopteran contaminant was ruled out. Since there were replicates of both taxa, it also seemed improbable that the differences were the result of a misidentification or a labeling error. The "Xiphocentronidae" came from the collection at the Smithsonian National Museum of Natural History curated by Dr. Oliver Flint Jr. The collection at the Smithsonian is organized by family, and alphabetically ordered. "Xiphocentronidae" is at the end, and when Kjer sampled the collection, he pulled legs from the undetermined caddisflies in the "X" case at the end. When we saw the unusual results from these taxa (Fig. 2), Frandsen wrote to Dr. Flint to see if he could help solve the mystery. He did. Flint had simply put these undetermined specimens in the last box where he knew he could find them. They were not xiphocentronids. Dr. Flint suggested that these taxa belong in the Sericostomatoidea so we performed a further analysis with the "mystery taxon" and other Sericostomatoidea. In this analysis, the "Xiphocentronidae sp." came out at sister to Beraeidae (Fig. 3). Through further communication, the identification was confirmed as Beraeidae similar to a species of *Ernodes* described by Nozaki & Kagaya (1994).

Through this analysis, we were able to place an enigmatic taxon, and fix an error in the barcode of life database for future identification accuracy. The other

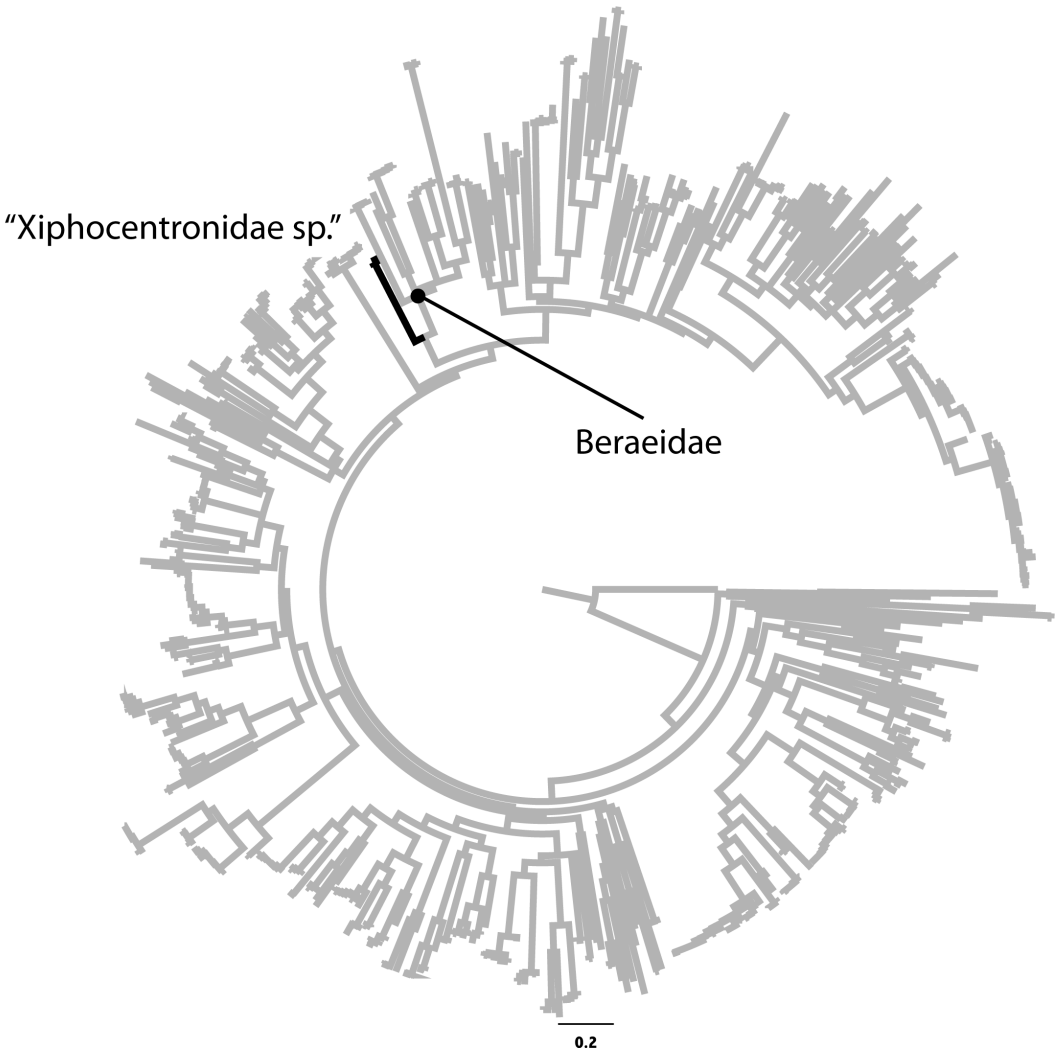

Figure 3. Phylogram generated in RAxML. Sericostomatoidea, dark branch represents the placement of "Xiphocentronidae sp."

problematic taxon in Fig. 2 was the New Zealand endemic, *Zelandoptila.* Since these appeared to be so different from the other xiphocentronids and psychomyiids, they were included in a larger analysis including representatives from all annulipalpian families. This analysis placed the *Zelandoptila* within Ecnomidae sister to *Daternomina* + *Ecnomina zealandica* (Fig. 4) which occur in nearby Australia in the case of *Daternomina* and New Zealand in the case of *Ecnomina zealandica*. This makes sense biogeographically as the only other member of Psychomyiidae in the Australian region is *Tinodes abberans* from New Guinea (Neboiss 1986). Upon the presentation of this material at the 14th International Symposium on Trichoptera, Dr. Flint also commented, anecdotally, that one of his past students had worked with the morphology of *Zelandoptila* and came to the same conclusion, that it belongs in Ecnomidae. Upon further investigation into the literature, Johanson & Espeland (2010) found a similar result that *Zelandoptila* was either a member of Ecnomidae or sister to Ecnomidae including *Pseudoneureclipsis*. Errors are inevitable in the barcode database, and arise from many possible sources. Here we demonstrate that barcodes are problematic without expert help in checking the voucher specimens. However, we also show that with thoughtful analysis, the problems can be solved.

After the problem taxa were removed, the question remains about the utility of our phylogeny of Xiphocentronidae plus Psychomyiidae (Fig. 1). There is much evidence that single gene phylogenies need not track species trees (Avise 1994),

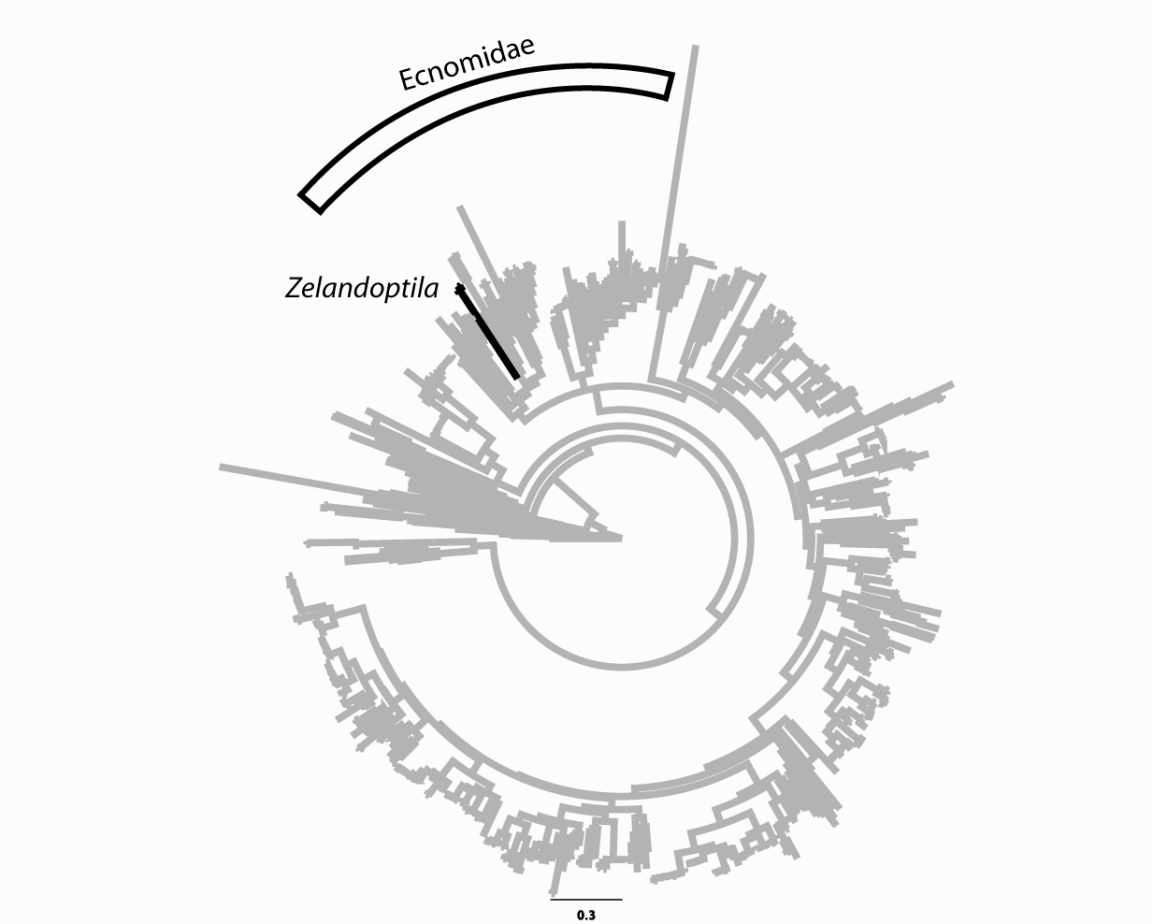

Fig. 4. Phylogram generated from RAxML for Psychomyioidea; *Zelandoptera* is represented by the dark branch nested within Ecnomidae.

because ancestral polymorphisms can sort independently of speciation. And Kjer *et al*. (2001) show that the COI is homoplastic for deep level Trichoptera phylogenetics. Where the problems begin to materialize is unknown. We are not saying that these results are ideal. But they do provide testable hypotheses for further analysis, and that is all any phylogeny can be. It seems a shame to us to have this mass of data available, and ignore what it has to say about phylogeny because we fear the answer may not be perfect. If we understand the limitations of the data, then we can take what we can from the phylogeny. For example, if one were to work on a generic revision, and there were no generic-level phylogenies available, we think that even a tentative but species-rich phylogeny from barcode data would be useful for selecting potential outgroups.

Although we take the moral "high ground" in stating that our phylogenies are merely "testable hypotheses," we may be on shaky philosophical grounds in selecting ingroups as close to the tips as possible, in order to mediate the problems with homoplasy in COI. The tree generated is based on the assumption that Xiphocentronidae plus Psychomyiidae is a "fact." Systematists often do this in their selection of ingroups and outgroups. If a phylogenetic analysis of Xiphocentronidae lacked representatives of Coleoptera, or even Echinodermata, there would be no objection because we have accepted that coleopterans and echinoderms are not trichopterans. Similarly, if we can corroborate the monophyly of Xiphocentronidae plus Psychomyiidae, as has been proposed by multiple studies (Frania & Wiggins 1997; Holzenthal *et al.*

2007; Kjer *et al.* 2001, 2002; Kjer *et al*., this symposium volume), as well as with our own data, then we should be safe to conduct an analysis of the two, with one family rooting the other. In addition, we have shown that our own analysis can locate those taxa that do not belong within the analysis, and recover a reasonable hypothesis that can be updated with additional data (Fig. 1).

These examples represent a few of the interesting outcomes of a phylogenetic analysis of the Trichoptera barcodes. In both of these cases, a molecular understanding of the specimens didn't replace the classical methods; rather, it enhanced and improved our understanding of the classification of these insects. We expect that further research using Trichoptera barcodes will yield similar new avenues to refine our taxonomic understanding of Trichoptera.

Literature Cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.

Anon (2010) *Adobe Illustator*. Adobe Systems, Inc., San Jose, California.

Avise, J.C. (1994) *Molecular Markers, Natural History and Evolution*. Springer.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792–1797.

Flint, O.S. & Kjer, K.M. (2011) A new species of *Neophylax* from northern Virginia, USA (Trichoptera: Uenoidae). *Proceedings of the Entomological Society of Washington*, 113, 7–13.

Frania, H.E. & Wiggins, G.B. (1997) Analysis of morphological and behavioural evidence for the phylogeny and higher classification of Trichoptera (Insecta). *Royal Ontario Museum*,*Life Sciences Contributions*, 160, 1–67.

Gouy, M., Guindon, S. & Gascuel, O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27, 221–224.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & de Waard, J.R. (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270, 313–321.

Hebert, P.D.N., Ratnasingham, S. & de Waard, J.R. (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, 270, S96–S99.

Holzenthal, R.W., Blahnik, R.J., Kjer, K.M. and Prather, A.L. (2007) An update on the phylogeny of caddisflies (Trichoptera). *In*: Bueno-Soria, J., Barba-Álvarez, R. & Armitage, B.J. (Eds.), *Proceedings of the XIIth International Symposium on Trichoptera*. The Caddis Press, Columbus, Ohio, pp. 143–153.

Johanson, K.A. & Espeland, M. (2010) Phylogeny of the Ecnomidae (Insecta: Trichoptera). *Cladistics*, 26, 36–48.

Kjer, K.M., Blahnik, R.J. & Holzenthal, R.W. (2001) Phylogeny of Trichoptera (Caddisflies): Characterization of signal and noise within multiple datasets. *Systematic Biology*, 50, 781–816.

Kjer, K.M., Blahnik, R.J. & Holzenthal, R.W. (2002) Phylogeny of caddisflies (Insecta, Trichoptera). *Zoologica Scripta*, 31, 83–91.

Kjer, K.M. & Honeycutt, R.L. (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evolutionary Biology*, 7–8.

Neboiss, A. (1986) *Atlas of Trichoptera of the SW Pacific - Australian Region*. 1st ed. Springer.

Nozaki, T. & Kagaya, T. (1994) A New *Ernodes* (Trichoptera, Beraeidae) from Japan. *Japanese Journal of Entomology*, 62, 193–200.

Rambaut, A. (2012) *FigTree Software*. Available from: http://tree.bio.ed.ac.uk/software/figtree/. (accessed 19 December 2014)

Resh, V.H. & Unzicker, J.D. (1975) Water quality monitoring and aquatic organisms: The importance of species identification. *Journal - Water Pollution Control Federation* 47, 9–19.

Ruiter, D.E., Boyle, E.E. & Zhou, X. (2013) DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. *BMC Ecology*, 13:5.

Stamatakis, A., Ludwig, T. & Meier, H. (2005) RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21, 456–463.

Sweeney, B.W., Battle, J.M., Jackson, J.K. & Dapkey, T. (2011) Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, 30, 195–216.

Swofford, D.L. (2003) *PAUP* Phylogenetic Analysis Using Parsimony* (**and Other Methods*). Sinauer Associates, Sunderland, Massachusetts.

Willis, J.C. & Yule, G.U. (1922) Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, 109, 177–179.

Zhou, X., Kjer, K.M. & Morse, J.C. (2007) Associating larvae and adults of Chinese Hydropsychidae caddisflies (Insecta:Trichoptera) using DNA sequences. *Journal of the North American Benthological Society*, 26, 719–742.

CHAPTER 2: AUTOMATIC SELECTION OF PARTITIONING SCHEMES FOR PHYLOGENETIC ANALYSES USING ITERATIVE *K*-MEANS CLUSTERING OF SITE RATES

Abstract

Background

Model selection is a vital part of most phylogenetic analyses, and accounting for the heterogeneity in evolutionary patterns across sites is particularly important. Mixture models and partitioning are commonly used to account for this variation, and partitioning is the most popular approach. Most current partitioning methods require some *a priori* partitioning scheme to be defined, typically guided by known structural features of the sequences, such as gene boundaries or codon positions. Recent evidence suggests that these *a priori* boundaries often fail to adequately account for variation in rates and patterns of evolution among sites. Furthermore, new phylogenomic datasets such as those assembled from ultra-conserved elements lack obvious structural features on which to define *a priori* partitioning schemes. The upshot is that, for many phylogenetic datasets, partitioned models of molecular evolution may be inadequate, thus limiting the accuracy of downstream phylogenetic analyses.

Results

We present a new algorithm that automatically selects a partitioning scheme via the iterative division of the alignment into subsets of similar sites based on their rates of evolution. We compare this method to existing approaches using a wide

range of empirical datasets, and show that it consistently leads to large increases in the fit of partitioned models of molecular evolution when measured using AICc and BIC scores. In doing so, we demonstrate that some related approaches to solving this problem may have been associated with a small but important bias.

Conclusions

Our method provides an alternative to traditional approaches to partitioning, such as dividing alignments by gene and codon position. Because our method is data-driven, it can be used to estimate partitioned models for all types of alignments, including those that are not amenable to traditional approaches to partitioning.

Background

The accuracy of phylogenetic inference often relies on the use of an appropriate model of molecular evolution (Sullivan and Joyce 2005; Steel 2005). Inaccurate tree reconstructions can be the result of both stochastic and systematic error. Stochastic error is the inevitable consequence of using finite datasets, and decreases as datasets grow in size. Systematic error results from biases such as the failure to adequately model the patterns of molecular evolution that generated the data (model misspecification) (Phillips, Delsuc, and Penny 2004; Felsenstein 2004; Yang and Rannala 2012), and can be amplified in large datasets, often resulting in strong support for the incorrect tree topologies

(Phillips, Delsuc, and Penny 2004; Felsenstein 1978; Jeffroy et al. 2006; Nishihara, Okada, and Hasegawa 2007; Rodríguez-Ezpeleta et al. 2007; Kumar et al. 2012). Improving approaches to model selection, even within existing phylogenetic frameworks, can help to reduce systematic error and improve the reliability of phylogenetic inference.

Accounting for the heterogeneity in the rates and patterns of evolution among sites in a DNA sequence alignment is an important part of selecting a model of molecular evolution (Yang 1996b; Buckley, Simon, and Chambers 2001; Lemmon and Moriarty 2004; Revell, Harmon, and Glor 2005; Bull et al. 1993). Among the methods proposed to account for this are mixture models (Pagel and Meade 2004; Le, Lartillot, and Gascuel 2008; Lartillot, Lepage, and Blanquart 2009) and partitioning (Nylander et al. 2004; Brandley, Schmitz, and Reeder 2005; Brown and Lemmon 2007; Kjer and Honeycutt 2007). Mixture models account for among-site heterogeneity by combining estimates of the likelihood of each site in the alignment under more than one model of molecular evolution. Partitioning accounts for among-site heterogeneity by splitting an alignment into several groups of sites (subsets) and estimating model parameters independently for each subset. Although mixture models are an elegant way to account for among-site heterogeneity, partitioning remains more popular, more widely implemented, and is currently the only approach that is computationally efficient enough to work on very large datasets (Stamatakis 2014; Guindon et al. 2010; Zwickl 2006; Ronquist et al. 2012; Drummond et al. 2012; Blair and

Murphy 2011; Lanfear et al. 2014b). Thus, our focus in this manuscript is on developing methods to improve the selection of partitioning schemes for phylogenetic analyses, with a view to improving the inference of phylogenetic trees from large datasets.

An inherent obstacle in partitioned phylogenetic analyses is the choice of an appropriate partitioning scheme. One approach would be to evaluate every possible partitioning scheme for a given dataset and choose the best scheme, perhaps according to one of the commonly used information theoretic metrics such as the AICc (Hurvich and Tsai 1989) or BIC (Schwarz 1978) or by some measure of biological features in the data. However, comparing all possible partitioning schemes is practically impossible because the number of partitioning schemes is astronomical even for very small alignments (Li, Lu, and Ortí 2008; Lanfear et al. 2012). For example, some of the smallest alignments used today, associated with DNA barcoding studies, contain ~658 base pairs (Hebert et al. 2003), which can be grouped into more than 1.0×10^{931} possible partitioning schemes: well beyond anything that can be feasibly analyzed by brute force. A related approach is to allow the data inform the assignment of sites to subsets, and to integrate out the uncertainty in these assignments in a Bayesian framework (Wu, Suchard, and Drummond 2013). Although this method is elegant, it has a high computational burden that renders it impractical for all but modestly sized datasets.

The most commonly used method for partitioning alignments, and the only one currently suited to very large datasets, is to define subsets according to structural features of the sequences in the alignment, such as gene boundaries, codon positions, structural components of rRNAs (such as stems and loops), or some combination of these. We call this 'traditional' partitioning throughout this manuscript. This approach is also known as mechanistic modeling because it describes known biological or mechanistic processes and is motivated by the biological observation that different molecular structural features can have different patterns of molecular evolution (Leavitt et al. 2013; Best and Stachowicz 2013; Springer et al. 1999; Brandley, Schmitz, and Reeder 2005; Brown and Lemmon 2007; E Biffin 2007; Bofkin and Goldman 2007; Liò and Goldman 1998; Hu, Shen, and Wang 2011; Huelsenbeck and Crandall 1997). Recently, various methods have been proposed to algorithmically refine traditional partitioning schemes by grouping together similar subsets of sites (Li, Lu, and Ortí 2008; Lanfear et al. 2012; Lanfear et al. 2014b). One example of this method is the PartitionFinder greedy algorithm (Lanfear et al. 2012), which works by joining a pre-defined subset with every other pre-defined subset and then selecting the grouping that most improves the AICc or BIC score. This is done iteratively until no more groupings improve the score. Using this method can result in large improvements in model fit. However, despite their popularity, all traditional partitioning approaches make an important assumption that is rarely questioned: that all of the sites in each of the pre-defined subsets (e.g. a

particular codon position in a particular gene) have evolved under a single evolutionary model.

A number of recent studies have suggested that traditional approaches to partitioning can be inadequate. Evidence suggests that there can be substantial heterogeneity of the evolutionary process within a single codon position of a single gene (Stergachis et al. 2013; Chris Simon et al. 1994; Chris Simon et al. 2006; Kjer and Honeycutt 2007; Yang 1996a; Lartillot and Philippe 2004) and within a single stem or loop of rRNA (Pagel and Meade 2004; Chris Simon et al. 2006; C Simon et al. 1996; Letsch and Kjer 2011). If this is true, then traditional approaches to partitioning may fail to adequately account for the variation in patterns of molecular evolution within each traditionally defined subset of sites. For smaller datasets, these limitations can be overcome by applying newer methods (Wu, Suchard, and Drummond 2013; Lartillot, Lepage, and Blanquart 2009), but for larger datasets the limitations of traditional partitioning remain a problem.

Another limitation of traditional partitioning involves its application to new types of molecular markers. Many of the latest methods for assembling phylogenomic datasets result in large alignments that consist either entirely or largely of non-protein coding DNA (e.g. introns and ultra-conserved elements (UCEs)) (Faircloth et al. 2012; Lemmon, Emme, and Lemmon 2012; McCormack et al. 2012; Crawford et al. 2012). It can be difficult to determine a good partitioning scheme

for these datasets with traditional approaches because we understand little about the molecular evolution of the sequenced regions, and the datasets lack convenient features such as codon positions on which subsets can be defined *a priori*. Thus, we face the problem that we lack adequate ways to model molecular evolution for some of the largest and most promising empirical datasets in our field.

One approach to choosing a partitioning scheme for large datasets is to group sites into subsets using estimates of site rates (Kjer and Honeycutt 2007; Kjer, Blahnik, and Holzenthal 2001; Ellingson et al. 2013; Cummins and McInerney 2011). Kjer and Honeycutt (Kjer and Honeycutt 2007) showed that partitioning an alignment in this way resulted in a mammal mitochondrial genome phylogeny that was better supported and more congruent with phylogenies based on nuclear data. Ellingson et al. (Ellingson et al. 2013) showed that this approach improved both topologies and node support for a phylogeny of fish. However despite their promise, these methods have not been widely adopted. This is perhaps because they are difficult to use and require various decisions (such as the appropriate number of subsets into which to divide the data) to be made before the analysis is conducted.

In this study, we develop a new algorithm that automatically defines partitioning schemes using site rates to cluster similar sites together into subsets. Our approach improves on previous work in three important ways. First, while

previous approaches (Kjer, Blahnik, and Holzenthal 2001; Kjer and Honeycutt 2007; Ellingson et al. 2013; Cummins and McInerney 2011; Misof et al. 2014) have required the user to choose the number of subsets before the analysis is carried out, our method estimates the optimal number of subsets directly from the data. This is important, because the optimal number of subsets may be difficult to predict in advance, and is influenced by several variables: e.g. the variation in substitution patterns among sites, the range of GTR submodels that can be selected for each subset, and by the method used to evaluate the fit of the model to the subset (e.g. AICc, BIC). Second, our method scales to work with the large datasets being produced today. Third, we explicitly test for, and address, the presence of a suspected bias in previous implementations of this approach: that the partitioning scheme selected by the method may be biased towards the phylogenetic tree from which the site-rates were calculated (Kjer, Blahnik, and Holzenthal 2001).

We demonstrate our approach on a wide range of datasets. Our results show that our method can be used to select partitioning schemes for the full range of datasets used in phylogenetics: from small barcoding datasets to large phylogenomic datasets consisting of ultra-conserved elements. In all cases, our method finds partitioning schemes that outperform those selected with traditional approaches to partitioning, when measured by metrics such as AICc and BIC.

Methods

Terminology

We follow the terminology established in other studies on partitioning (Lanfear et al. 2012; Li, Lu, and Ortí 2008; Lanfear et al. 2014b) in which a 'subset' refers to a set of sites for which parameters of a nucleotide substitution model will be independently estimated. Each site can only be assigned to one subset. In the phylogenetics community, a subset is often referred to as a 'partition'; we avoid using the word 'partition' because it has conflicting definitions in other fields (Lanfear et al. 2014b; Lanfear et al. 2012). A 'partitioning scheme' constitutes a collection of subsets that include every site in the alignment once and only once.

Iterative k-means partitioning algorithm

We present an algorithm (Fig. 1) that automatically selects a partitioning scheme for a given alignment without the need for pre-defined subsets. We first give an overview, and then expand on each step below:

1. Estimate a starting tree topology from the multiple sequence alignment;
2. Start with a partitioning scheme that has all sites assigned to a single subset, and choose the best-fit substitution model for that subset;
3. Calculate the information theoretic score of the current partitioning scheme;
4. For each subset in the current partitioning scheme, test whether that subset should be further divided:
    a. Generate site rates for the focal subset;

b. Divide the focal subset into two subsets using *k*-means clustering;
c. Choose the best-fit substitution model for each of the two new subsets;
d. Calculate the information-theoretic score of the partitioning scheme in which the two new subsets from 4c replace the focal subset;
e. If the information theoretic score improves, label the focal subset for division.

5. If no subsets have been labeled for division, terminate the algorithm. Otherwise, define a new partitioning scheme in which each labeled subset in the list is replaced by the two correspondingly smaller subsets defined in step 4b and return to step 4.

In step 1, we estimate a tree topology with branch lengths for the dataset. All likelihood calculations require a topology with branch lengths and a substitution model. Optimizing the tree topology at each step would be computationally intensive, particularly for large phylogenomic datasets. For this reason, we use a fixed tree topology throughout the course of the algorithm. In principle, any method to estimate a starting tree could be used since it has been argued that a non-random tree is likely to be sufficient for model selection (Abdo et al. 2005; Posada and Crandall 2001; Minin et al. 2003); in our implementation of the algorithm, we use the BioNJ algorithm implemented in PhyML (Guindon et al.

2010) to estimate a neighbor joining starting tree, then re-optimize the branch lengths of this tree in PhyML using the GTR+I+G model.

In step 2, we define a partitioning scheme in which all sites in the alignment are assigned to a single subset, and we then select a best-fit model of molecular evolution for this subset. The model selection step uses an information theoretic metric (e.g. the AICc or the BIC) to choose a substitution model from a list of candidate models. Here we select the best model from the set of 56 submodels of the GTR model available in PartitionFinder v1.1.1 (Lanfear et al. 2014b). These include the GTR and some of the most popular submodels implemented in PhyML, along with the model extensions using discrete GAMMA distributed site rates (+G) and/or a proportion of invariant sites (+I). During the model selection step, PartitionFinder provides two options for estimating branch lengths: ‘linked’ or ‘unlinked’. When the branch lengths are ‘unlinked’, all branch lengths are re-estimated for each model in the list. When branch lengths are ‘linked’, the relative branch lengths are determined by the tree estimated in step 1, and each model is afforded a single rate multiplier, which can stretch or shrink all branch lengths in tandem. Although ‘unlinked’ branch lengths allow users to better account for heterotachy (variation in relative branch lengths among subsets), in practice, they add so many parameters to the overall substitution that they are rarely preferred. For that reason, in what follows, we use ‘linked’ branch lengths in all of our analyses, although the option to use ‘unlinked’ branch lengths remains.

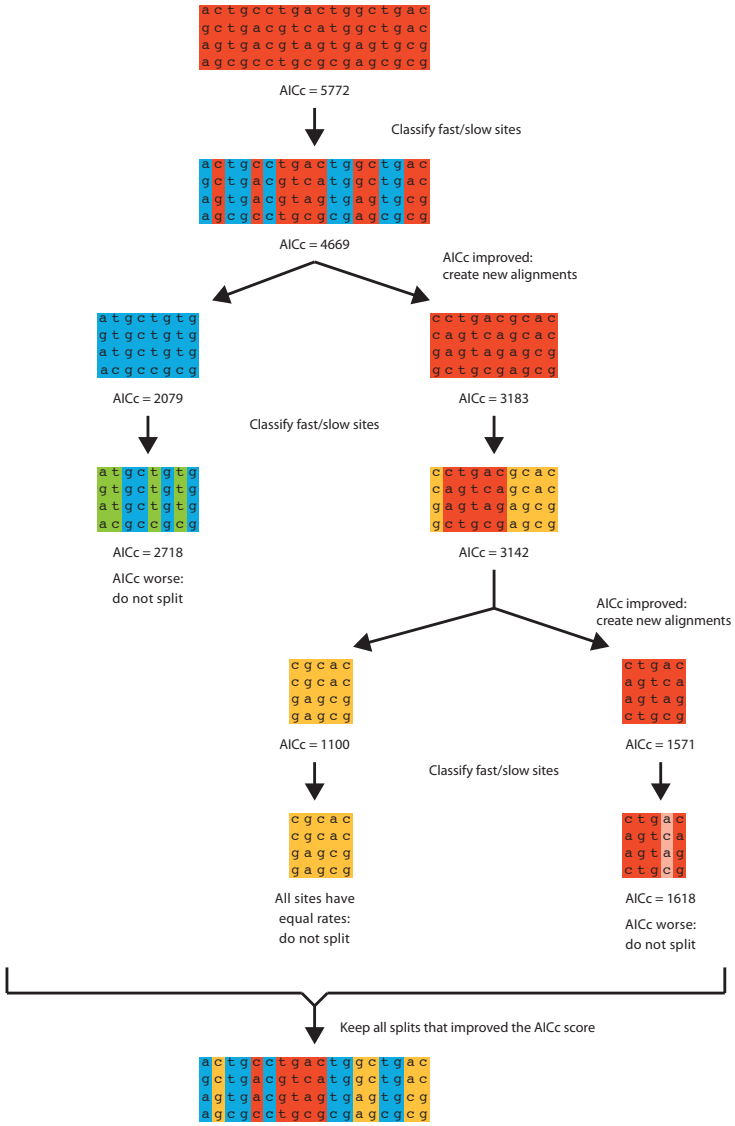

Figure 1. This figure illustrates the progress of a hypothetical run of the iterative *k*-means algorithm. The algorithm commences with an alignment that is treated as a single subset, and for which the AICc score has been calculated (step 3 in the description in the main text; represented by the red sequence alignment at the top). During this step, each of 56 GTR+I+G submodels is fit to the alignment and the model that returns the best AICc score is chosen. Next, the algorithm calculates TIGER site rates for each site (step 4a in the description in the main text), and uses these rates to classify the sites of the alignment into fast (red) and slow (blue) sites using the *k*-means algorithm (step 4b in the description in the main text). The AICc score of a model in which these two subsets are treated independently is then calculated (steps 4c-d in the description in the main text). If the score improves, the split is accepted. The fast (red) and slow (blue) sites are then used to create two new alignments, and the process is repeated with each new subset. This continues until no more subset splits are accepted. The final step combines all splits that improved the AICc score to create a single partitioning scheme for the dataset.

In step 3 we calculate one of two information theoretic scores (the AICc and the BIC (Lanfear et al. 2012)) of the current partitioning scheme. At the start of the algorithm, when all sites are assigned to a single subset, this score is equal to that of the initial best-fit substitution model in our iterative algorithm.

In step 4, we decide whether to subdivide each of the subsets in the current partitioning scheme. In step 4a, we fit a GTR+GAMMA model of molecular evolution to the subset, conditioned on the tree and its relative branch lengths estimated during step 1 using maximum likelihood in PhyML (Guindon et al. 2010). Then we use one of two methods to calculate site-specific rates for each site in the subset, (i) likelihood-based site rates or (ii) Tree Independent Generation of Evolutionary Rates (TIGER) site rates (Cummins and McInerney 2011). Likelihood site-rates are estimated in PhyML with GTR+G and using the "--print_site_lnl" option. Likelihood site-rates depend on the branch lengths and therefore have to be recomputed for subsets that have been divided. Similarly, a changing composition of site patterns in subsets as they are divided requires the TIGER rates to be recomputed since they depend on other sites in the subset alignment. TIGER site rates are calculated using a non-tree based method that estimates the similarity among site patterns as a surrogate for evolutionary rates (Cummins and McInerney 2011). This method relies on the construction and comparison of set partitions for each alignment pattern. For example, if a given alignment pattern is "AACGGA", the resulting set partition

would be $P(i)$ = {{1, 2, 6}, {3}, {4, 5}}. We call $P(i)$ the "character partition" of site *i*. $P(i)$ consists of a set of at most four sets, that contain the sequence numbers in the alignment pattern that have, respectively, the nucleotides A, C, G, or T at site *i.* The number of non empty sets, which we denote by $|P(i)|$ is equal to the number of different nucleotides found in site pattern *i*. The character partition of each site is then compared to the character partition of every other site. The sites are evaluated for agreement with every other site using a "partition agreement score", $(pa(i,j))$, which is defined as:

$$pa(i,j) = \frac{\sum_{x \in P(j)} a(x, P(i))}{|P(j)|}$$

where $a\big(x, P(i)\big)$ is equal to 0 or 1 depending on whether *x* is compatible with the character partition of site *i*, i.e. if *x* is a subset of one of the sets in $P(i)$:

$$a\big(x, P(i)\big) = \begin{cases} 1 \; if \; x \subseteq A \; for \; some \; A \in P(i) \\ 0 \end{cases}$$

The rate (r_i) for the alignment pattern at site *i* is then obtained by computing the mean partition agreement score across all sites:

$$r_i = \frac{\sum_{j \neq i} pa(i,j)}{n-1}$$

where n is equal to the number of sites in the alignment. The sites that are more similar to the pool of sites in the alignment are considered slow with rates approaching 1.0 (invariant sites always return a rate of 1.0), while the sites that are less similar to the pool of sites in the alignment are considered fast with rates approaching 0. It is important to note that this method ignores the character state in an alignment pattern when the set partitions are compared,

e.g. the set partition and resulting site rate of “AACGGA” would be identical to that of “TTGAAT”. Although software exists to calculate TIGER site rates (Cummins and McInerney 2011), we found the existing implementation to be too slow to be useful. Instead, TIGER site rates are calculated using a fast, C++ based program that we developed (Paul Frandsen and Christoph Mayer, n.d.).

In step 4b, we use the *k*-means clustering algorithm to divide the sites in the focal subset into two clusters based on one or more of the site-wise parameter estimates from step 4a. *K*-means is a fast clustering algorithm capable of handling large datasets with high dimensionality (MacQueen 1967; Lloyd 1982). It clusters data points by minimizing the within-cluster sum of squares measured between each data point and its closest cluster ‘centroid’. The goal of *k*-means is to minimize the function:

$$\min_{\{\mu_1,\ldots,\mu_k\}} \sum_{h=1}^{k} \sum_{x \in \chi_h} ||x - \mu_h||^2$$

Where *k* is the number of clusters, μ is the cluster centroid, and *x* is any given data point, in the case of this study, the site rates, and $||x||$ is the L_2 norm, or Euclidean length, of x. The algorithm proceeds through two steps:

1. The assignment step, in which each point is assigned to a cluster with its closest centroid.
2. The update step, in which cluster centroids are moved to the center (mean) of their new clusters.

The number of clusters (*k*) is chosen *a priori* and fixed at 2 in our case, and then *k* centroids are placed within the sample space. The initial placement of centroids is an important step; poor placement can result in an unsatisfactory exploration of the sample space and, although the algorithm may converge, it may only reach a local optimum. To avoid this, we use the *k*-means++ centroid initialization method, which has been shown to be superior when compared to other centroid seeding techniques such as random placement (Ostrovsky et al. 2006; Arthur and Vassilvitskii, n.d.). We perform 100 initializations of the *k*-means algorithm, selecting the initialization that best minimizes the within-cluster sum of squares. Following initialization, Euclidean distances between each data point and the centroids are calculated and each data point is assigned to a cluster based on its nearest centroid. The centroids are then moved to the mean of their respective clusters (the *k*-mean) and distances are recalculated. This process is repeated until the centroids no longer move beyond a threshold at the end of the iteration. We used the *k*-means algorithm from the scikit-learn package implemented in Python (Pedregosa et al. 2012). In theory, any statistic that can be estimated on a site-specific basis could be used for clustering. In what follows, we compare the performance of likelihood site rates and TIGER site rates.

In steps 4c and 4d, we use the output of the *k*-means algorithm to create two new subsets, and then use an information-theoretic metric to decide whether splitting the focal subset improves the overall model of molecular evolution. To do that, we first (step 4c) estimate the best model for each of the two new subsets from our set of candidate models as described above. We then (step 4d) calculate the information-theoretic score of two partitioning schemes: one in which the focal subset is retained as a single subset, and one in which the focal subset is divided into two new subsets. If the overall information theoretic score of the latter partitioning scheme is better, we label the focal subset as one that should be divided.

Once step 4 has been applied to all of the subsets in the current partitioning scheme, we ask whether there are any subset divisions that improved the overall information theoretic score (step 5). If there are none, then the algorithm terminates, since we are unable to find a partitioning scheme better than the current scheme. Otherwise, we divide all of the subsets that are labeled for division in step 4. Then the algorithm iterates.

Empirical considerations

The algorithm above makes the assumption that likelihoods can be calculated for any collection of sites in an alignment. During the development of the algorithm, we found some cases in which PhyML was unable to analyze some subsets. This was usually because the alignments were too small or contained

only sites with identical site patterns. Since our aim is to produce partitioning schemes that can be used to estimate phylogenetic trees with programs like PhyML, and since these problematic subsets are likely to occur during any approach similar to the one we describe here, we designed the following solution. First, we flag the problematic subsets as the algorithm proceeds, and make the conservative assumption that their site-likelihoods will be identical to their site-likelihoods in the larger subset from which they were generated. This allows us to estimate conservative information theoretic scores for partitioning schemes as the algorithm proceeds. At the end of the algorithm (i.e. after step 5), we combine each of the problematic subsets with their nearest neighbor subset, defined as the non-problematic subset with the centroid (estimated in step 3a) that has the shortest Euclidean distance to the centroid of the problematic subset. This process is repeated until there are no problematic subsets, i.e. until PhyML can successfully analyze all of the subsets in the partitioning scheme.

Empirical evaluation

To evaluate the performance of the iterative *k*-means algorithm, we compared ten partitioning scheme selection approaches on ten different datasets (Table 1). The approaches comprise five different partitioning methods, each of which was applied with both the BIC and AICc (Table 2). The five methods we compared were: (i) no partitioning (i.e. treating all sites as belonging to a single subset); (ii) partitioning by gene and codon position/rDNA stems and loops (all); (iii)

Table 1. Names, references, and clade information for the datasets used in empirical analyses

Dataset Name	Clade (latin)	Clade (common)	Paper Reference	Dataset Reference
Anderson 2014	Cephalopoda: Loliginidae	pencil squids	(Anderson et al. 2014)	(Anderson et al. 2013)
Cognato 2001	Coleoptera: Scolytinae	bark beetles	(Cognato and Vogler 2001b)	(Cognato and Vogler 2001a)
Grande 2013	Paracanthopterygii	paracanthopterygian fish	(T. Grande 2013)	(Grande, Borden, and Smith 2013)
Kang 2013a	Xiphophorus	swordtail fish	(Kang et al. 2013)	N/A
Kawahara 2013	Hyposmocoma	Hawaiian fancy-cased caterpillar	(A Y Kawahara and Rubinoff 2013)	(Akito Y. Kawahara and Rubinoff 2013)
Kjer 2007	Mammalia	mammals	(Kjer and Honeycutt 2007)	N/A
Leavitt 2013	Acridoidea	grasshoppers	(Leavitt et al. 2013)	N/A
McCormack 2013	Neoaves	birds	(McCormack et al. 2013a)	(McCormack et al. 2013b)
Oaks 2011	Crocodylia	crocodilians	(Oaks 2011b)	(Oaks 2011a)
Sharanowski 2011	Braconidae	parisitoid wasps	(Sharanowski, Dowling, and Sharkey 2011b)	(Sharanowski, Dowling, and Sharkey 2011a)

optimizing the partitioning scheme from (ii) using the greedy algorithm implemented in PartitionFinder 1.1.1; (iv) iterative *k*-means with likelihood site-rates; (v) iterative *k*-means with TIGER site-rates.

During the empirical evaluation, one dataset, McCormack 2013 (McCormack et al. 2013a), was too large and partitioned into too many pre-defined subsets to analyze with PartitionFinder's greedy algorithm in a reasonable amount of time. For this dataset, we used the relaxed clustering algorithm (Lanfear et al. 2014a) in PartitionFinder 1.1.1. Relaxed clustering is optimized for large datasets and uses RAxML (Stamatakis 2006) for all likelihood calculations. Since only two nucleotide substitution models are implemented in RAxML (GTR+G and GTR+I+G) we used a two-step approach. First, the optimal partitioning scheme was selected using the relaxed clustering algorithm for the two RAxML models, and second, we reselected models for each subset of the initial partitioning scheme with the 'user' option in PartitionFinder 1.1.1, but this time with PhyML and considering the full set of models used in every other treatment. This allowed us to directly compare the information theoretic scores of this partitioning scheme with those selected by the other methods.

Starting tree bias evaluation

Although it has been shown that a starting tree topology is unlikely to negatively affect model selection as long as it is non-random (Posada and Crandall 2001), it was unclear whether this was true when using the iterative *k*-means method

we develop here. Specifically, we were unsure whether site rates calculated under the assumption that the starting tree is true would bias our partitioning schemes toward recovering the starting tree during downstream phylogenetic analyses. Thus, we designed a simple test to assess whether this introduced a bias.

To test whether the starting tree introduced bias into the estimation of the partitioning scheme, we used a five-step process. First, we estimated a neighbor-joining (NJ) tree for the data. Second, we created twenty new trees, where each new tree was a single subtree-prune and regraft (SPR) move away from the NJ tree, giving a set of 20 plausible non-random trees for the dataset. Third, we used these 20 trees as starting trees from which we estimated 20 partitioning schemes for each dataset using three methods: the PartitionFinder greedy algorithm, iterative *k*-means with likelihood site-rates, and iterative *k*-means with TIGER site rates (i.e. 60 partitioning schemes in total, for each dataset). Fourth, we estimated a maximum-likelihood phylogenetic tree using RAxML for all 60 partitioning schemes for each dataset. For each of the three methods we compared, the process resulted in a collection of 20 distinct starting trees, and 20 estimated ML trees. The final step in the process involved statistically testing whether the starting trees are more similar to their corresponding ML trees than would be expected by chance. To do this, we used a bootstrap test in which the observed test statistic is the sum of the Robinson-Foulds (Robinson and Foulds 1981) distances between each starting

tree and the corresponding ML tree (i.e. the ML tree estimated from the partitioning scheme that assumed the corresponding starting tree). For example, in the most extreme case, where each ML tree is identical to its corresponding starting tree, the observed test statistic would be zero. The null distribution of this test statistic is then estimated by re-calculating the test-statistic 999 times after randomly shuffling the list of ML trees each time. If the starting tree biases the estimation of the ML tree, then we expect the observed test statistic to be in the lower tail of the null distribution. We calculate the one-tailed p-value from the position of the observed test statistic in a ranked list of the values of the test statistic from the null distribution.

Simulation example

While the primary purpose of this paper was to evaluate the efficacy of the iterative *k*-means algorithm on empirical datasets by comparing the relative fit of each model using information theoretic metrics like AICc and BIC, we also evaluated our method with a simple simulation. First, we simulated a tree under the Yule (pure-birth) process in INDELible v1.03 (Fletcher and Yang 2009). We chose a rooted tree and specified the following parameters for the simulation: number of tips-100, birth-0.1, and death-0 with a tree depth of 0.1. We then simulated a 1,000 bp alignment using the Jukes Cantor (Jukes and Cantor 1969) model. Next, we scaled the tree from the first run to a tree depth of 1.0 and simulated another 1,000 bp alignment using Jukes Cantor. Finally, we concatenated the alignments (total: 2,000 bp) and estimated a partitioning

scheme for it using iterative *k*-means with TIGER rates. Each step, from tree simulation through partitioning scheme selection was repeated 20 times. These conditions were chosen to explicitly test whether the iterative *k*-means algorithm would 1) assign alignment sites to subsets containing other sites generated from the same model, and 2) find the correct number of subsets.

Results and Discussion

The primary purpose of this study was to evaluate the performance of the iterative *k*-means algorithm on empirical data and compare those results to other commonly used partitioning strategies. To do this, we selected partitioning schemes for existing empirical datasets using several different methods and compared the relative fit of each partitioning scheme using AICc and BIC.

The iterative *k*-means algorithm substantially outperformed all other partitioning approaches for each of the ten datasets we analyzed, regardless of the details of the *k*-means approach or the information theoretic metric we used (Table 2, Figures 2 and 3). The set of alignments that we used to test the algorithm comprised a wide range of lengths, number of taxa, and types of molecular markers, confirming the utility of our new algorithm for a wide range of phylogenetic analyses.

Figures 2 and 3 show comparisons of the AICc and BIC scores achieved by five partitioning methods: using a single partition; partitioning according to structural

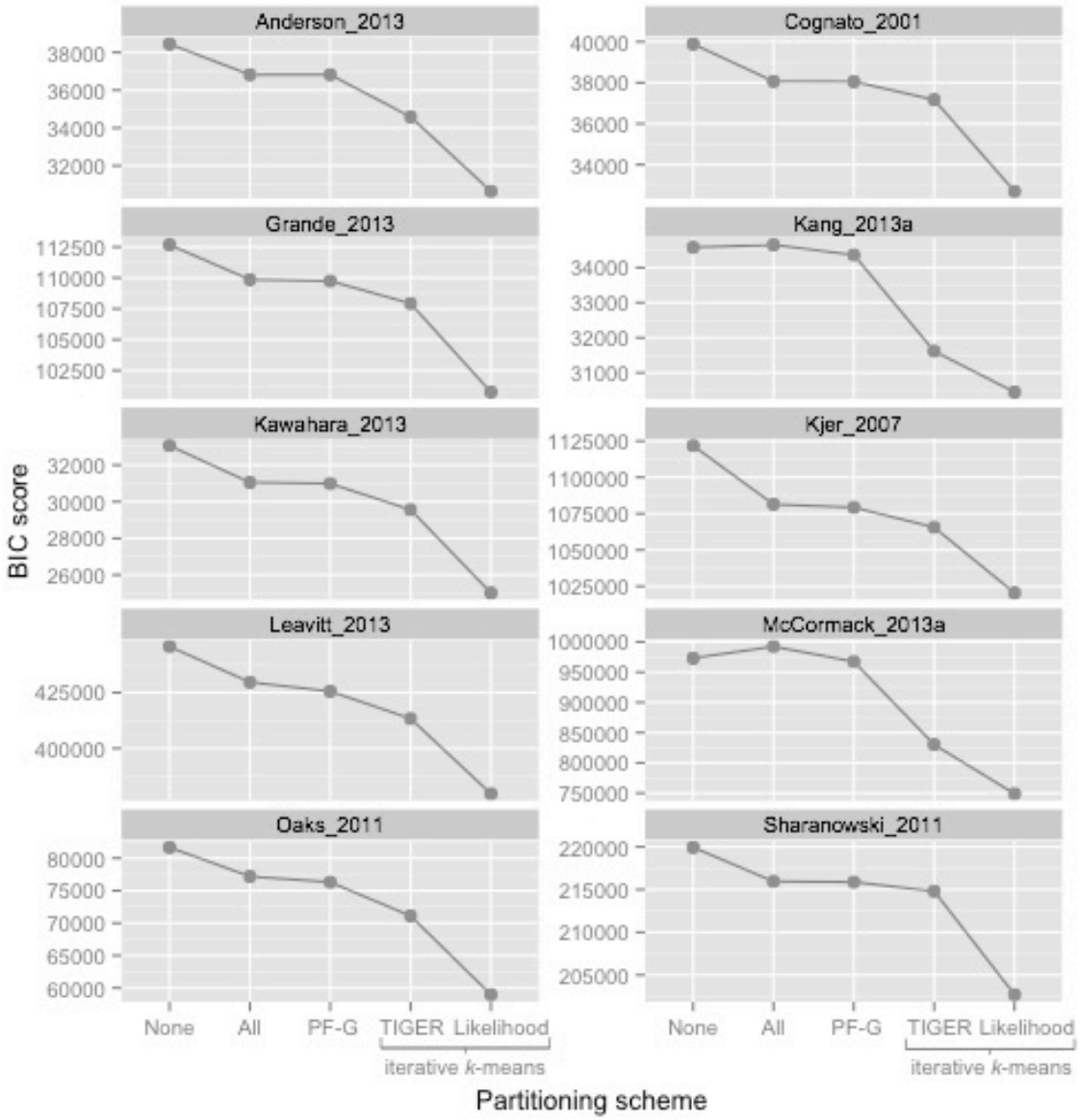

Figure 2. BIC scores for partitioning schemes estimated during empirical testing (lower is better). The *k*-means methods presented here outperform traditional methods. “None” is no partitioning, “All” is the user partitioning scheme, “PF-G” is the PartitionFinder greedy algorithm, “TIGER” is iterative *k*-means using TIGER site rates, “Likelihood” is iterative *k*-means using likelihood site rates. Note: The “PF-G” score for the McCormack 2013 dataset was obtained using the PartitionFinder relaxed clustering followed by model selection with PhyML as described in the methods, not the greedy algorithm.

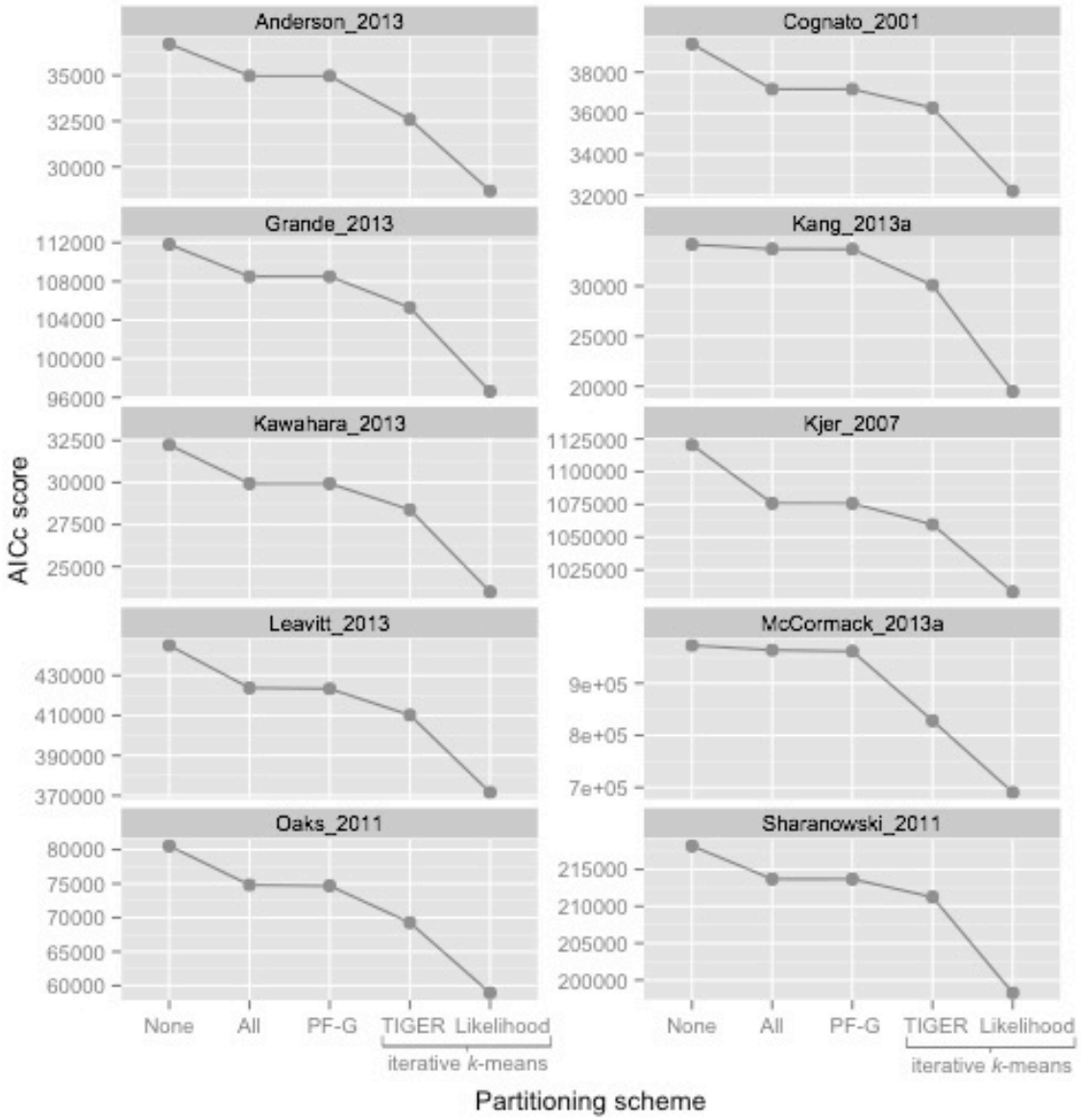

Figure 3. AICc scores for partitioning schemes estimated during empirical testing. (lower is better). The *k*-means methods presented here outperform traditional methods. “None” is no partitioning, “All” is the user partitioning scheme, “PF-G” is the PartitionFinder greedy algorithm, “TIGER” is iterative *k*-means using TIGER site rates, “Likelihood” is iterative *k*-means using likelihood site rates. Note: The “PF” score for the McCormack 2013 dataset used the PartitionFinder relaxed clustering followed by model selection with PhyML as described in the methods, not the greedy algorithm.

Table 2. AICc and BIC scores for every partitioning scheme selected for each dataset.

Dataset Name	no partitioning	user	greedy	TIGER site rates	likelihood site rates
			AICc		
Anderson 2014	36721	34972	34972	32584	28681
Cognato 2001	39373	37172	37172	36262	32216
Grande 2013	111819	108501	108501	105300	96616
Kang 2013	34157	33699	33682	30091	19490
Kawahara 2013	32238	29940	29940	28385	23488
Kjer 2007	1120392	1075872	1075749	1059594	1008180
Leavitt 2013	445052	423826	423389	410220	371585
McCormack 2013	972461	963551	961143	828180	690645
Oaks 2011	80550	74814	74686	69278	58947
Sharanowski 2011	218188	213696	213696	211251	198268
			BIC		
Anderson 2014	38444	36820	36820	34590	30649
Cognato 2001	39890	38067	38056	37168	32701
Grande 2013	112673	109862	109741	107929	100751
Kang 2013	34578	34650	34364	31617	30448
Kawahara 2013	33057	31052	30989	29561	25020
Kjer 2007	1121793	1081392	1079270	1065998	1020447
Leavitt 2013	445625	429638	425574	413343	379692
McCormack 2013	973122	992389	967479	830140	748946
Oaks 2011	81673	77190	76316	71117	58988
Sharanowski 2011	219937	215953	215872	214802	202707

features of the sequences; optimizing a partitioning scheme based on structural features using PartitionFinder; iterative *k*-means partitioning with likelihood-based site rates; and iterative *k*-means partitioning based on site rates estimated using the TIGER method. Figures 2 and 3 show that both *k*-means methods we describe here consistently outperform all of the other methods. The figures also suggest that the likelihood-based method is superior, as it consistently outperforms the method based on TIGER rates, achieving lower AICc and BIC scores. However, the apparent superiority of the likelihood-based method comes at a cost – it is also frequently associated with a bias: phylogenetic trees estimated from partitioning schemes derived from the likelihood-based approach were often more similar to the starting trees than would be expected by chance (Table 3). In 4 out of 9 datasets (Table 3), our test for starting tree bias returned a statistically significant result for the likelihood-based method.

In contrast, when using the TIGER based rates we found no evidence for starting tree bias in any of the datasets that we examined. We attribute the difference between these two methods to the fact that the likelihood-based approach relies on a particular starting tree to calculate rates of evolution, whereas the TIGER method calculates rates without assuming a particular tree (Cummins and McInerney 2011). It appears that the dramatic gains in AICc and BIC scores achieved using the likelihood-based *k*-means approach are partially attributable to overfitting the partitioning scheme to the starting tree, and that

Table 3. P-values and effect sizes for each dataset from starting tree bias analysis. Analyses with p-values of less than .05 were found to have significant starting tree bias.

Dataset	Partitioning method	p-value	effect size
Anderson_2013	greedy algorithm	1	0
Anderson_2013	lnL rates *k*-means	0.03	-0.604
Anderson_2013	TIGER rates *k*-means	1	0
Cognato_2001	greedy algorithm	1	0
Cognato_2001	lnL rates *k*-means	0.006	-0.424
Cognato_2001	TIGER rates *k*-means	0.202	-0.152
Grande_2013	greedy algorithm	1	0
Grande_2013	lnL rates *k*-means	0.047	-0.225
Grande_2013	TIGER rates *k*-means	1	0
Kang_2013a	greedy algorithm	1	0
Kang_2013a	lnL rates *k*-means	0.391	-0.105
Kang_2013a	TIGER rates *k*-means	1	0.14
Kawahara_2013	greedy algorithm	0.397	-0.095
Kawahara_2013	lnL rates *k*-means	0.008	-0.348
Kawahara_2013	TIGER rates *k*-means	0.828	0.051
Leavitt_2013	greedy algorithm	1	0
Leavitt_2013	lnL rates *k*-means	1	0
Leavitt_2013	TIGER rates *k*-means	1	0
McCormack_2013a	greedy algorithm	0.409	-0.116
McCormack_2013a	lnL rates *k*-means	0.52	-0.048
McCormack_2013a	TIGER rates *k*-means	0.158	-0.084
Oaks_2011	greedy algorithm	1	0.094
Oaks_2011	lnL rates *k*-means	1	0.019
Oaks_2011	TIGER rates *k*-means	1	0
Sharanowski_2011	greedy algorithm	1	0
Sharanowski_2011	lnL rates *k*-means	0.056	-0.304
Sharanowski_2011	TIGER rates *k*-means	0.069	-0.222

this overfitting can then bias subsequent phylogenetic analyses. One symptom of this overfitting is that the likelihood-based rates method often selected subsets of sites that consisted entirely of invariant sites of a single nucleotide state. Such subsets are difficult if not impossible to justify on biological grounds. Together, these characteristics suggest that the likelihood method is problematic, and should be avoided. For the remainder of the paper, we focus only on the results from our study that used rates calculated with the TIGER method, which do not show these undesirable characteristics.

One of the primary motivations for this study was to develop a method to select partitioning schemes for datasets that are very large and/or that comprise molecular markers that are not amenable to traditional partitioning approaches, both of which are increasingly common (Faircloth et al. 2012; McCormack et al. 2012; Crawford et al. 2012). It is encouraging, therefore, to note that the iterative *k*-means algorithm performed particularly well on the phylogenomic bird dataset (Table 1, Figures 2 and 3) (McCormack et al. 2013a), which was both very large and comprised solely of UCE's, for which traditional approaches to partitioning are difficult to apply. For example, when each UCE was placed in its own subset, the BIC score was worse than when all UCE's were grouped into a single subset (BIC scores of 992,389 and 973,121 respectively (Table 2)). When the partitioning scheme was selected using the relaxed clustering algorithm in PartitionFinder, the BIC score improved to 967,478 (Fig. 2, Table 2), but when using the iterative *k*-means method with TIGER rates, the BIC score improved to

830,140. This represents a substantial improvement to the partitioning schemes selected using traditional methods (Fig. 2, Table 2).

The iterative *k*-means clustering also worked well for datasets consisting of protein coding genes from the standard phylogenetic toolbox. A close examination of the partitioning schemes reveals that the algorithm chooses subsets that reflect the traditional biological partitioning boundaries such as genes and codon positions (Figs. 4-5). For example, in the partitioning schemes selected for the Hawaiian fancy-case caterpillar dataset consisting of three genes (A Y Kawahara and Rubinoff 2013), the k-means approach resulted in one large subset that contained almost all first and second codon position sites across all three genes along with some third codon position sites (Fig. 4, subset 1, Fig. 5), and eleven smaller subsets which consisted primarily of third codon positions sites from the three loci (Fig. 4, subsets 2-12, Fig. 5). Insofar as it broadly combines first and second codon positions, and separates out third codon positions, this partitioning scheme is similar to a popular traditional approach to partitioning that does the same (Shapiro, Rambaut, and Drummond 2006). Although some of the structure of the classical partitioning boundaries exists in the subsets chosen by *k*-means, other subsets include sites from a wide range of genes and codon positions (Fig. 4 subsets 1-4, 6, 8). These results confirm that there is biological value to partitioning by genes and codon positions, but also suggest that relying solely on such boundaries may often fail

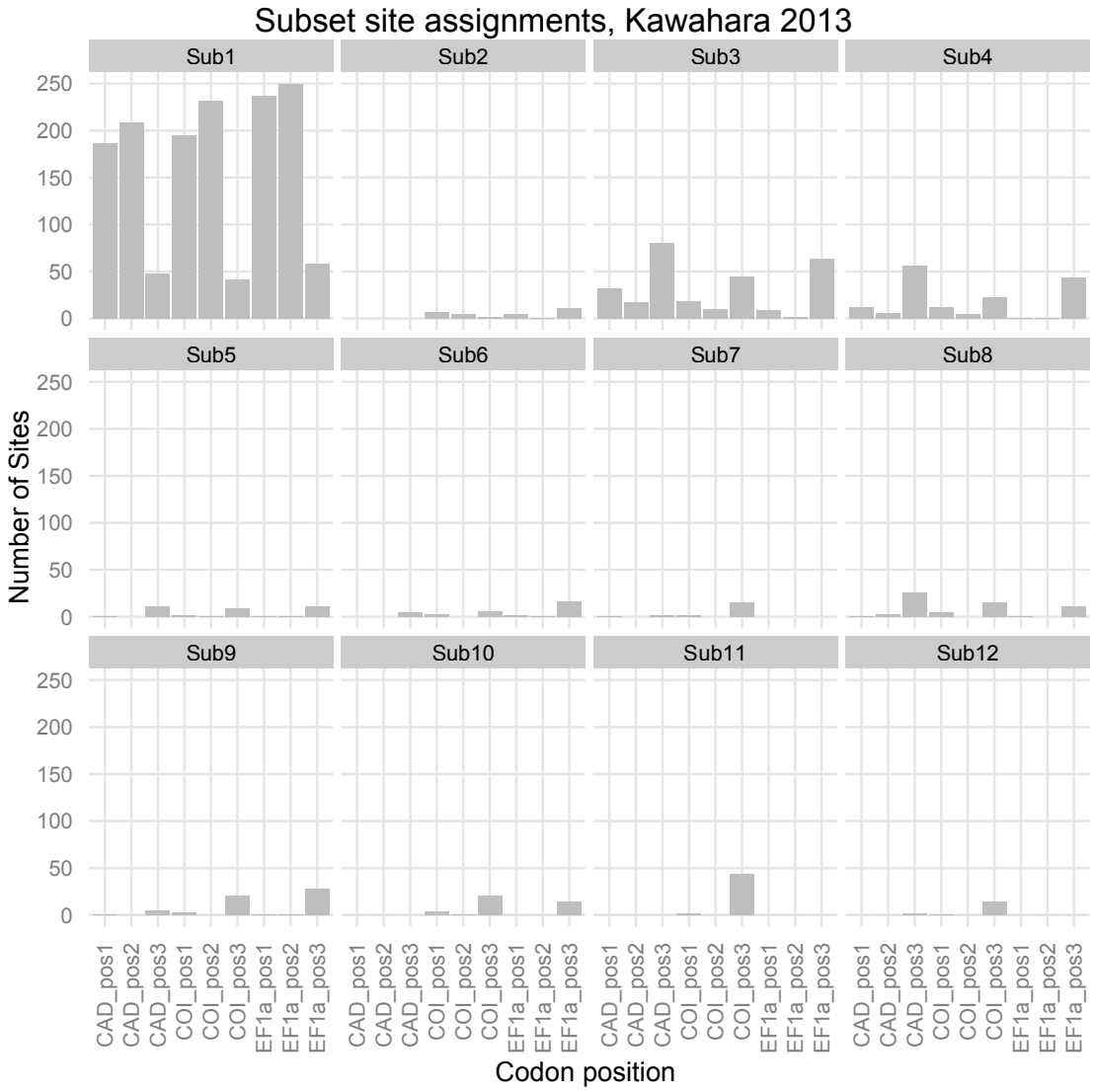

Figure 4. Assignment of codon position by gene to subsets selected using the iterative *k*-means algorithm clustered using TIGER site rate estimates on the Kawahara 2013 dataset. Subsets are ordered by the mean site rate from slowest to fastest. Sites from each codon position are spread throughout the subsets with the majority of variation among sites in the 3rd codon position.

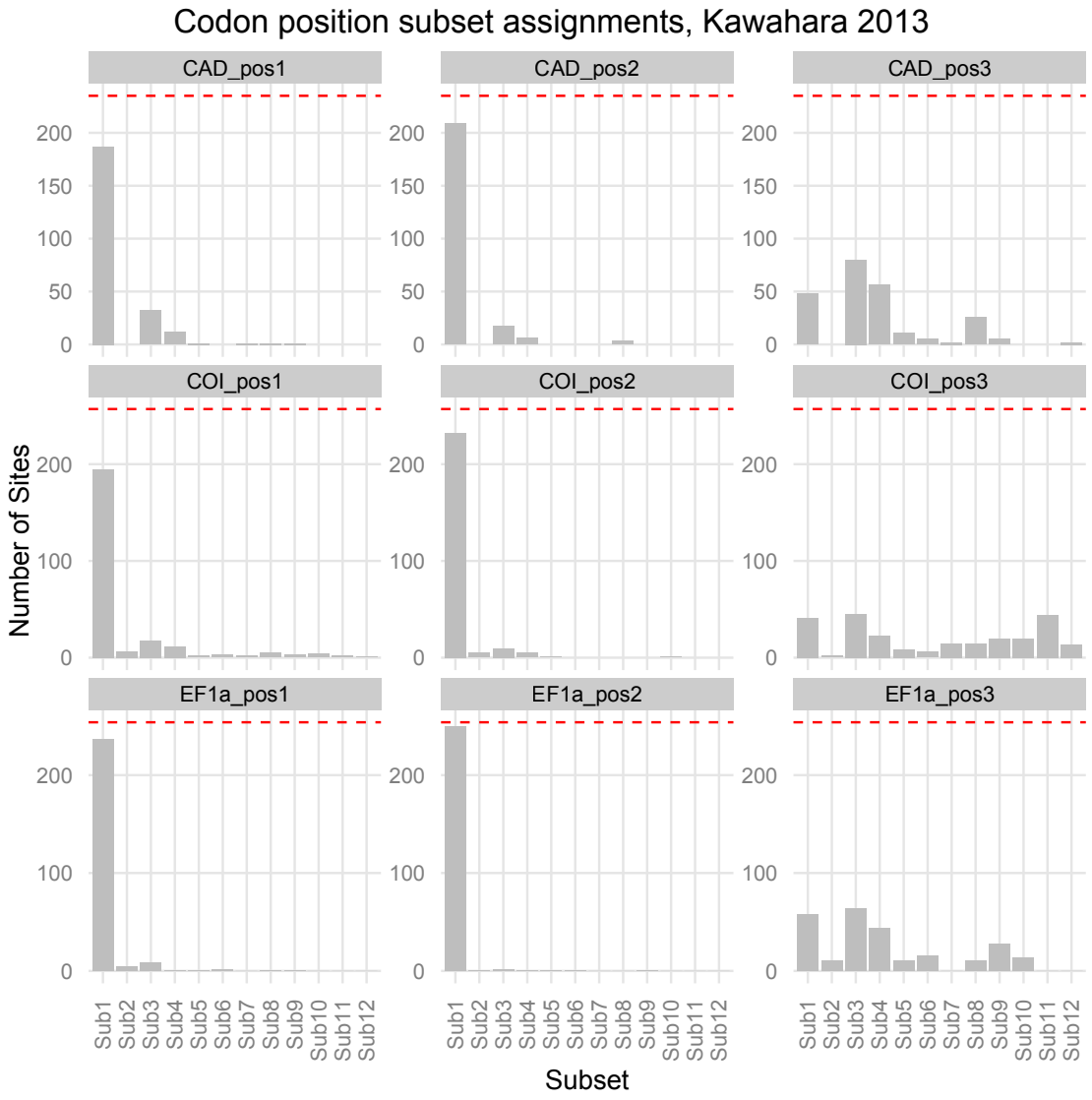

Figure 5. Subset assignments for the sites from each codon position using the iterative *k*-means algorithm clustering using TIGER site rate estimates on the Kawahara 2013 dataset. Each row corresponds to a single gene and each column corresponds to a different codon position. The dotted red line represents the total number of sites in each codon position. In each chart, subsets are ordered by the mean site rate from slowest to fastest. First and second codon positions most closely align with "traditional partitioning", while substantial variation exists among the 3rd codon position sites.

to capture some of the complex patterns of molecular evolution among sites, potentially limiting the accuracy of downstream phylogenetic analyses.

The iterative k-means algorithm provides a powerful data-driven method to account for complex patterns of variation in rates of molecular evolution among sites. This is primarily because it tends to group together sites that evolve at similar rates of evolution, reducing the need for additional parameters to describe variation in rates across sites within a given group of sites (Fig. 6). For example, in the crocodilian dataset [71], although 15 subsets were selected in the partitioning scheme chosen with TIGER rates for a dataset with just over 7,000 sites, models with the GAMMA model of rate heterogeneity were never chosen, and the proportion of invariable sites model was chosen for only three subsets. In contrast, the partitioning scheme chosen with the greedy algorithm included 11 subsets with seven that chose GAMMA or proportion of invariable sites for among site rate variation correction. Out of 168 total subsets selected using iterative *k*-means with TIGER rates and evaluated with BIC during our empirical evaluation, 77 (45.8%) required the additional parameters of gamma, proportion of invariable sites, or both. In contrast, of the 92 subsets chosen with the PartitionFinder greedy algorithm, 86 (93.5%) of the models included gamma, proportion of invariable sites, or both. These results support recent observations that more flexible models of variation in rates among sites tend to fit the data much better than those that rely on distributional assumptions (Soubrier et al. 2012; Galtier et al. 2006), and suggest that the iterative *k*-means approach to

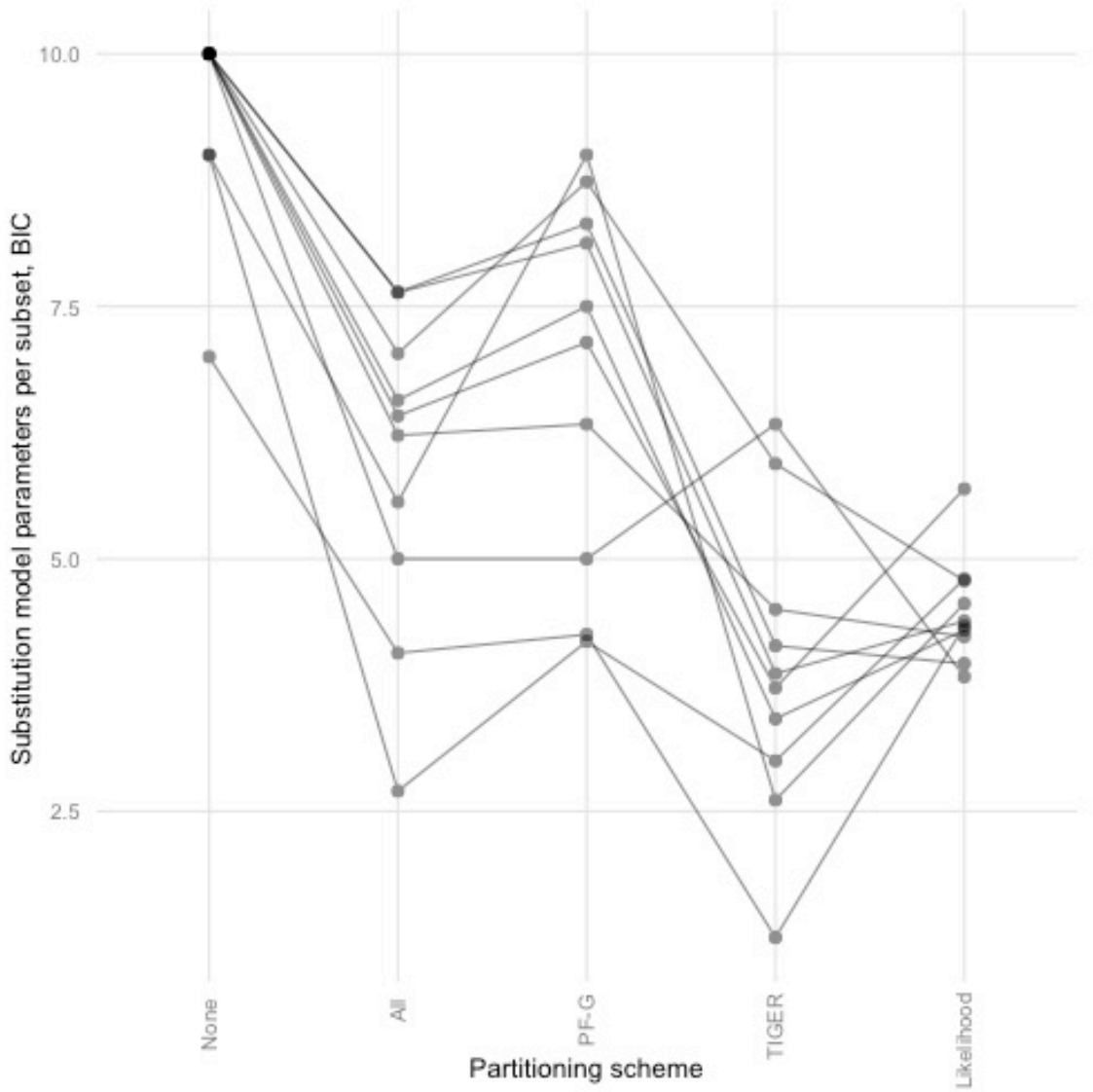

Figure 6. Average number of parameters per subset for different partitioning scheme estimation methods using BIC. Each line represents a different empirical dataset. “None” is no partitioning, “All” is the user partitioning scheme, “PF-G” is the PartitionFinder greedy algorithm, “TIGER” is iterative *k*-means using TIGER site rates, “Likelihood” is iterative *k*-means using likelihood site rates. The parameters per subset decrease for the *k*-means methods.

partitioning may be particularly useful when the variation in rates across sites cannot be adequately modeled using a combination of traditional partitioning (e.g. using genes and codon positions) and gamma-distributed rates (Kjer and Honeycutt 2007; Galtier et al. 2006). Methods for accounting for this kind of heterogeneity exist and include the CAT model, implemented into the program Phylobayes (Lartillot, Lepage, and Blanquart 2009; Lartillot and Philippe 2004; Lartillot and Philippe 2006; Lartillot, Brinkmann, and Philippe 2007; Quang, Gascuel, and Lartillot 2008) and a "spike and slab" model recently described by Wu et al. (Wu, Suchard, and Drummond 2013) that has been implemented into BEAST 2 (Bouckaert et al. 2014). Our method provides a scalable, maximum likelihood based alternative to these approaches.

We evaluated the partitioning schemes chosen by the iterative *k*-means with TIGER rates for the simulated alignments based on the criteria that, 1) alignment sites generated under the same model would be assigned to the same subsets, and 2) the correct number of subsets would be chosen. Our results show that most subsets consisted primarily of sites generated from the same model (Fig. 7). For example, when we searched for subsets that consisted of 95% or more sites generated under the same model, we found 263 of 289 (91%) that met the 95% cutoff (Table 4). However, the number of subsets varied from 12-24 (Table 4), far more than the two subsets under which the data were simulated. To further understand this behavior, we examined the partitioning schemes generated after each iteration of the algorithm. We found that the first split often

Table 4. The number of subsets selected using the iterative *k*-means algorithm for 20 simulated alignments in which 2 independent subsets were simulated.

Simulation Replicate	*k*-means subsets	Number of subsets consisting of >=95% sites from same model
1	14	12
2	12	12
3	16	15
4	14	13
5	13	12
6	16	14
7	15	15
8	13	12
9	14	13
10	13	12
11	14	12
12	15	13
13	14	12
14	12	11
15	14	13
16	15	13
17	14	14
18	24	20
19	13	12
20	14	13

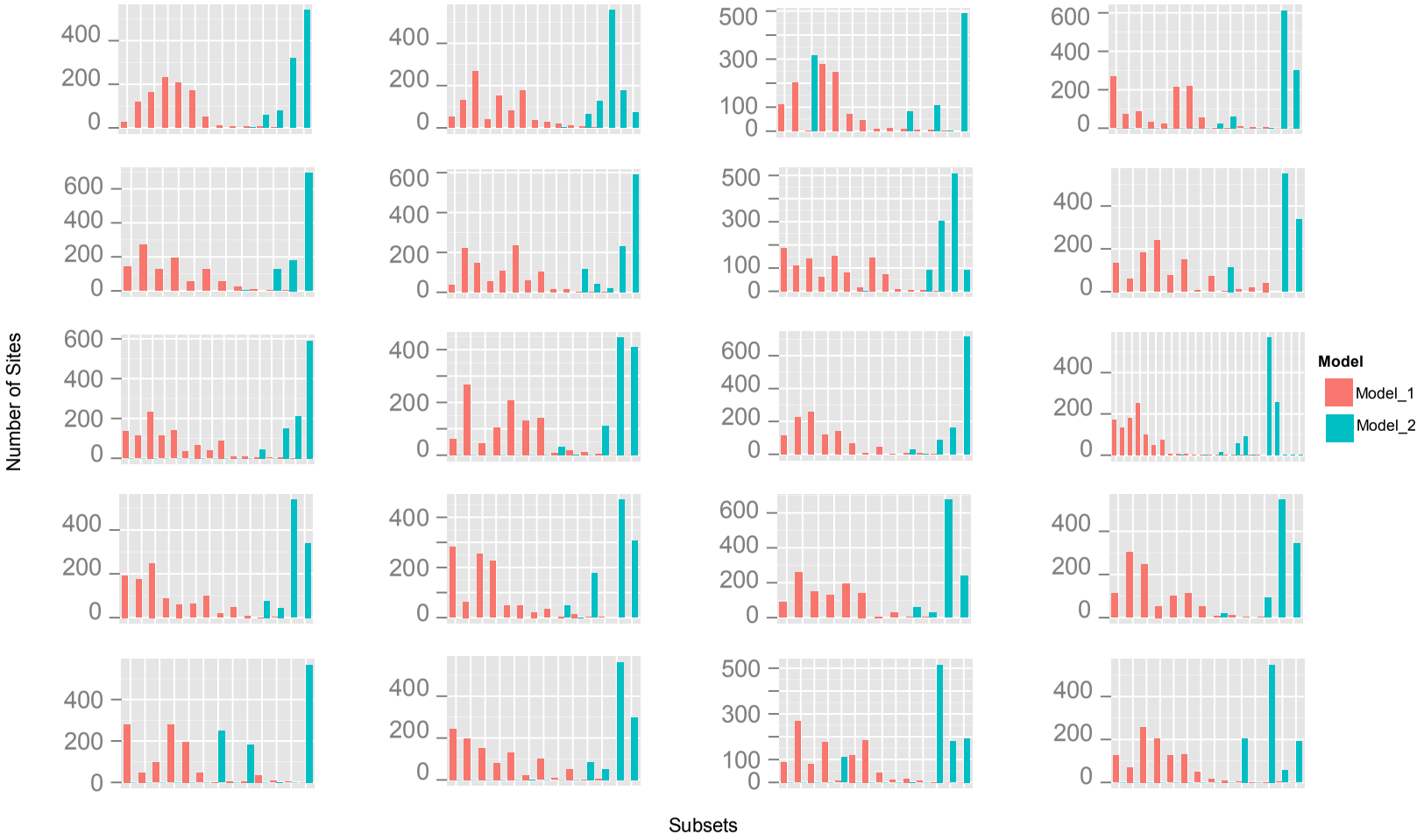

Figure 7. Assignment of sites using iterative *k*-means and TIGER site rates for 20 simulated alignments. The different colors represent sites generated under different models. The number of subsets selected is variable while sites are most often clustered with other sites simulated under the same model.

closely approximated the true model, but due to continual increases in the BIC score, many more splits were accepted. This result suggests that the inability to recover the true number of subsets could be due to the nature of the metrics for the evaluation of model fit. Whatever the underlying reason for the over-partitioning of simulated datasets, these results suggest that when using methods like these to select partitioning schemes for empirical studies, it would be prudent to estimate phylogenetic trees under a range of intermediate partitioning schemes as well as the final partitioning scheme. An important next step in investigating these and other approaches to partitioning is a full-scale simulation study which examines a broad range of simulation conditions, and which assesses the effects of each not only on recovering the correct model, but also on recovering the correct tree.

Despite the failure of the k-means method to recover the correct number of subsets in simulated data, three factors suggest that this finding is unlikely to severely compromise the method. First, previous studies have shown that defining too many partitions may have negligible impact on downstream phylogenetic inferences such as tree topologies, bootstrap support, or branch lengths (Brown and Lemmon 2007; Li, Lu, and Ortí 2008). Second, on empirical datasets, the *k*-means method tends to select a relatively modest number of subsets – never more than double the number of features in the dataset itself (e.g. individual codon positions in individual genes), and often many fewer. For example, for the McCormack et al. dataset (McCormack et al. 2013a;

McCormack et al. 2013b), there were 416 individual UCE's, and the k-means method selected just 18 subsets of sites. Third, the *k*-means method selects partitioning schemes that make biological sense with respect to what we already know about variation in rates and patterns of evolution (Figs. 4-5).

It is important to note that the iterative *k*-means algorithm represents a heuristic search for an optimal partitioning scheme. As such, it cannot be guaranteed to find the optimum partitioning scheme for any given dataset. Furthermore, the *k*-means algorithm itself is somewhat stochastic in nature, and so it is likely that repeated analyses of the same dataset might lead to the estimation of partitioning schemes with very minor differences. Although we have focused on DNA sequence alignments in this study, the approach we describe can also be applied to amino acid alignments.

Our research suggests that the iterative *k*-means algorithm is an improvement over traditional approaches to partitioning. Accounting for variation of rates among sites has long been viewed as a vital part of modeling in phylogenetics (Yang 1994; Yang 1996b; Kjer, Blahnik, and Holzenthal 2001; Kjer and Honeycutt 2007; Chris Simon et al. 1994; Chris Simon et al. 2006; Lartillot and Philippe 2004), and we have shown that using site rates to inform subset assignments results in substantial improvements in the AICc and BIC scores of partitioning schemes, when compared to more commonly used methods. Perhaps most importantly, the iterative *k*-means algorithm provides a data

driven method for modeling patterns of molecular evolution in markers such as UCE's that have been difficult to model with traditional approaches.

Conclusion

Partitioning remains the most commonly used method for accounting for variation in the rates and patterns of molecular evolution among sites in phylogenetic analyses. As the size and number of phylogenomic datasets grows, it is increasingly important to fit more realistic partitioned models to those datasets. The algorithm we present in this paper does this by automatically selecting a partitioning scheme for datasets of variable size and type without the need of an *a priori* determination of partition boundaries or number of desired subsets. Although we identified potential pitfalls of using such algorithms (such as a starting tree bias when using likelihood site rates), we also showed how these pitfalls could be overcome. These methods provide an important step forward in improving our approaches to modeling molecular evolution, particularly for very large datasets, as well as suggesting fruitful directions for further improvements.

References

Abdo, Zaid, Vladimir N. Minin, Paul Joyce, and Jack Sullivan. 2005. "Accounting for Uncertainty in the Tree Topology Has Little Effect on the Decision-Theoretic Approach to Model Selection in Phylogeny Estimation." *Molecular Biology and Evolution* 22 (3): 691–703. doi:10.1093/molbev/msi050.

Anderson, Frank E., Alexis Bergman, Samantha H. Cheng, M. Sabrina Pankey, and Tooraj Valinassab. 2014. "Lights out: The Evolution of Bacterial

Bioluminescence in Loliginidae." *Hydrobiologia* 725 (1): 189–203. doi:10.1007/s10750-013-1599-1.

Anderson, Frank E., Alexis Bergman, Samantha H. Cheng, M. Sabrina Pankey, Tooraj Valinassab, and Frank E. Anderson. 2013. "Data from: Lights out: The Evolution of Bacterial Bioluminescence in Loliginidae." *Dryad Digital Repository*, June. http://hdl.handle.net/10255/dryad.47872.

Arthur, David, and Sergei Vassilvitskii. n.d. "K-Means++: The Advantages of Careful Seeding."

Best, Rebecca J., and John J. Stachowicz. 2013. "Phylogeny as a Proxy for Ecology in Seagrass Amphipods: Which Traits Are Most Conserved?" *PLoS ONE* 8 (3): e57550. doi:10.1371/journal.pone.0057550.

Blair, Christopher, and Robert W. Murphy. 2011. "Recent Trends in Molecular Phylogenetic Analysis: Where to Next?" *Journal of Heredity* 102 (1): 130–38. doi:10.1093/jhered/esq092.

Bofkin, Lee, and Nick Goldman. 2007. "Variation in Evolutionary Processes at Different Codon Positions." *Molecular Biology and Evolution* 24 (2): 513–21. doi:10.1093/molbev/msl178.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Comput Biol* 10 (4): e1003537. doi:10.1371/journal.pcbi.1003537.

Brandley, Matthew C., Andreas Schmitz, and Tod W. Reeder. 2005. "Partitioned Bayesian Analyses, Partition Choice, and the Phylogenetic Relationships of Scincid Lizards." *Systematic Biology* 54 (3): 373–90. doi:10.1080/10635150590946808.

Brown, Jeremy M., and Alan R. Lemmon. 2007. "The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics." *Systematic Biology* 56 (4): 643–55. doi:10.1080/10635150701546249.

Buckley, Thomas R., Chris Simon, and Geoffrey K. Chambers. 2001. "Exploring Among-Site Rate Variation Models in a Maximum Likelihood Framework Using Empirical Data: Effects of Model Assumptions on Estimates of Topology, Branch Lengths, and Bootstrap Support." *Systematic Biology* 50 (1): 67–86. doi:10.1080/10635150116786.

Bull, J. J., John P. Huelsenbeck, Clifford W. Cunningham, David L. Swofford, and Peter J. Waddell. 1993. "Partitioning and Combining Data in Phylogenetic Analysis." *Systematic Biology* 42 (3): 384–97. doi:10.1093/sysbio/42.3.384.

Cognato, Anthony I., and Alfried P. Vogler. 2001a. "Data from: Exploring Data Interaction and Nucleotide Alignment in a Multiple Gene Analysis of Ips (Coleoptera: Scolytinae)." *Dryad Digital Repository*. http://hdl.handle.net/10255/dryad.678.

———. 2001b. "Exploring Data Interaction and Nucleotide Alignment in a Multiple Gene Analysis of Ips (Coleoptera: Scolytinae)." *Systematic Biology* 50 (6): 758–80. doi:10.1080/106351501753462803.

Crawford, Nicholas G., Brant C. Faircloth, John E. McCormack, Robb T. Brumfield, Kevin Winker, and Travis C. Glenn. 2012. “More than 1000 Ultraconserved Elements Provide Evidence That Turtles Are the Sister Group of Archosaurs.” *Biology Letters* 8 (5): 783–86. doi:10.1098/rsbl.2012.0331.

Cummins, Carla A., and James O. McInerney. 2011. “A Method for Inferring the Rate of Evolution of Homologous Characters That Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases.” *Systematic Biology* 60 (6): 833–44. doi:10.1093/sysbio/syr064.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. “Bayesian Phylogenetics with BEAUti and the BEAST 1.7.” *Molecular Biology and Evolution* 29 (8): 1969–73. doi:10.1093/molbev/mss075.

E Biffin, M. G. Harrington. 2007. “Structural Partitioning, Paired-Sites Models and Evolution of the ITS Transcript in Syzygium and Myrtaceae.” *Molecular Phylogenetics and Evolution* 43 (1): 124–39. doi:10.1016/j.ympev.2006.08.013.

Ellingson, Ryan A., Camm C. Swift, Llyod T. Findley, and David K. Jacobs. 2013. “Convergent Evolution of Ecomorphological Adaptations in Geographically Isolated Bay Gobies (Teleostei: Gobionellidae) of the Temperate North Pacific.” *Molecular Phylogenetics and Evolution*. doi:10.1016/j.ympev.2013.10.009.

Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales.” *Systematic Biology* 61 (5): 717–26. doi:10.1093/sysbio/sys004.

Felsenstein, Joseph. 1978. “Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.” *Systematic Zoology* 27 (4): 401–10. doi:10.2307/2412923.

———. 2004. *Inferring Phylogenies*. Sunderland, Mass.: Sinauer Associates.

Fletcher, William, and Ziheng Yang. 2009. “INDELible: A Flexible Simulator of Biological Sequence Evolution.” *Molecular Biology and Evolution* 26 (8): 1879–88. doi:10.1093/molbev/msp098.

Galtier, Nicolas, David Enard, Yoan Radondy, Eric Bazin, and Khalid Belkhir. 2006. “Mutation Hot Spots in Mammalian Mitochondrial DNA.” *Genome Research* 16 (2): 215–22. doi:10.1101/gr.4305906.

Grande, Terry, W. Calvin Borden, and William Leo Smith. 2013. “Data from: Limits and Relationships of Paracanthopterygii: A Molecular Framework for Evaluating Past Morphological Hypotheses.” *Dryad Digital Repository*. http://hdl.handle.net/10255/dryad.41087.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance

of PhyML 3.0." *Systematic Biology* 59 (3): 307–21. doi:10.1093/sysbio/syq010.

Hebert, Paul D N, Alina Cywinska, Shelley L Ball, and Jeremy R deWaard. 2003. "Biological Identifications through DNA Barcodes." *Proceedings of the Royal Society B: Biological Sciences* 270 (1512): 313–21. doi:10.1098/rspb.2002.2218.

Huelsenbeck, John P., and Keith A. Crandall. 1997. "Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood." *Annual Review of Ecology and Systematics* 28 (1): 437–66. doi:10.1146/annurev.ecolsys.28.1.437.

Hu, Gang, Shiyi Shen, and Kui Wang. 2011. "On the Evolution Rate in Mammalian Mitochondrial Genomes." *Computational Biology and Chemistry* 35 (3): 137–42. doi:10.1016/j.compbiolchem.2011.04.001.

Hurvich, Clifford M., and Chih-Ling Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76 (2): 297–307. doi:10.1093/biomet/76.2.297.

Jeffroy, Olivier, Henner Brinkmann, Frédéric Delsuc, and Hervé Philippe. 2006. "Phylogenomics: The Beginning of Incongruence?" *Trends in Genetics* 22 (4): 225–31. doi:10.1016/j.tig.2006.02.003.

Jukes, TH, and CR Cantor. 1969. "Evolution of Protein Molecules." In *Mammalian Protein Metabolism*. Academy Press.

Kang, Ji Hyoun, Manfred Schartl, Ronald B. Walter, and Axel Meyer. 2013. "Comprehensive Phylogenetic Analysis of All Species of Swordtails and Platies (Pisces: Genus Xiphophorus) Uncovers a Hybrid Origin of a Swordtail Fish, Xiphophorus Monticolus, and Demonstrates That the Sexually Selected Sword Originated in the Ancestral Lineage of the Genus, but Was Lost Again Secondarily." *BMC Evolutionary Biology* 13 (1): 25. doi:10.1186/1471-2148-13-25.

Kawahara, Akito Y., and Daniel Rubinoff. 2013. "Data from: Convergent Evolution in the Explosive Hawaiian Fancy Cased Caterpillar Radiation." *Dryad Digital Repository*, July. http://hdl.handle.net/10255/dryad.48986.

Kawahara, A Y, and D Rubinoff. 2013. "Convergent Evolution of Morphology and Habitat Use in the Explosive Hawaiian Fancy Case Caterpillar Radiation." *Journal of Evolutionary Biology* 26 (8): 1763–73. doi:10.1111/jeb.12176.

Kjer, Karl M., Roger J. Blahnik, and Ralph W. Holzenthal. 2001. "Phylogeny of Trichoptera (Caddisflies): Characterization of Signal and Noise Within Multiple Datasets." *Systematic Biology* 50 (6): 781–816. doi:10.1080/106351501753462812.

Kjer, Karl M, and Rodney L Honeycutt. 2007. "Site Specific Rates of Mitochondrial Genomes and the Phylogeny of Eutheria." *BMC Evolutionary Biology* 7 (1): 8. doi:10.1186/1471-2148-7-8.

Kumar, Sudhir, Alan J. Filipski, Fabia U. Battistuzzi, Sergei L. Kosakovsky Pond, and Koichiro Tamura. 2012. "Statistics and Truth in Phylogenomics."

Molecular Biology and Evolution 29 (2): 457–72. doi:10.1093/molbev/msr202.

Lanfear, Robert, Brett Calcott, Simon Y. W. Ho, and Stephane Guindon. 2012. "PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses." *Molecular Biology and Evolution* 29 (6): 1695–1701. doi:10.1093/molbev/mss020.

Lanfear, Robert, Brett Calcott, David Kainer, Christoph Mayer, and Alexandros Stamatakis. 2014a. "Selecting Optimal Partitioning Schemes for Phylogenomic Datasets." *BMC Bioinformatics* in press.

———. 2014b. "Selecting Optimal Partitioning Schemes for Phylogenomic Datasets." *BMC Evolutionary Biology* 14 (1): 82. doi:10.1186/1471-2148-14-82.

Lartillot, Nicolas, Henner Brinkmann, and Hervé Philippe. 2007. "Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model." *BMC Evolutionary Biology* 7 (Suppl 1): S4. doi:10.1186/1471-2148-7-S1-S4.

Lartillot, Nicolas, Thomas Lepage, and Samuel Blanquart. 2009. "PhyloBayes 3: A Bayesian Software Package for Phylogenetic Reconstruction and Molecular Dating." *Bioinformatics (Oxford, England)* 25 (17): 2286–88. doi:10.1093/bioinformatics/btp368.

Lartillot, Nicolas, and Hervé Philippe. 2004. "A Bayesian Mixture Model for across-Site Heterogeneities in the Amino-Acid Replacement Process." *Molecular Biology and Evolution* 21 (6): 1095–1109. doi:10.1093/molbev/msh112.

———. 2006. "Computing Bayes Factors Using Thermodynamic Integration." *Systematic Biology* 55 (2): 195–207. doi:10.1080/10635150500433722.

Leavitt, James R., Kevin D. Hiatt, Michael F. Whiting, and Hojun Song. 2013. "Searching for the Optimal Data Partitioning Strategy in Mitochondrial Phylogenomics: A Phylogeny of Acridoidea (Insecta: Orthoptera: Caelifera) as a Case Study." *Molecular Phylogenetics and Evolution* 67 (2): 494–508. doi:10.1016/j.ympev.2013.02.019.

Lemmon, Alan R., Sandra A. Emme, and Emily Moriarty Lemmon. 2012. "Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics." *Systematic Biology* 61 (5): 727–44. doi:10.1093/sysbio/sys049.

Lemmon, Alan R., and Emily C. Moriarty. 2004. "The Importance of Proper Model Assumption in Bayesian Phylogenetics." *Systematic Biology* 53 (2): 265–77. doi:10.1080/10635150490423520.

Le, Si Quang, Nicolas Lartillot, and Olivier Gascuel. 2008. "Phylogenetic Mixture Models for Proteins." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1512): 3965–76. doi:10.1098/rstb.2008.0180.

Letsch, Harald O., and Karl M. Kjer. 2011. "Potential Pitfalls of Modelling Ribosomal RNA Data in Phylogenetic Tree Reconstruction: Evidence from Case Studies in the Metazoa." *BMC Evolutionary Biology* 11 (1): 146. doi:10.1186/1471-2148-11-146.

Li, Chenhong, Guoqing Lu, and Guillermo Ortí. 2008. “Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci.” *Systematic Biology* 57 (4): 519–39. doi:10.1080/10635150802206883.

Liò, Pietro, and Nick Goldman. 1998. “Models of Molecular Evolution and Phylogeny.” *Genome Research* 8 (12): 1233–44. doi:10.1101/gr.8.12.1233.

Lloyd, S. 1982. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory* 28 (2): 129–37. doi:10.1109/TIT.1982.1056489.

MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In . The Regents of the University of California. http://projecteuclid.org/euclid.bsmsp/1200512992.

McCormack, John E., Brant C. Faircloth, Nicholas G. Crawford, Patricia Adair Gowaty, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Are Novel Phylogenomic Markers That Resolve Placental Mammal Phylogeny When Combined with Species-Tree Analysis.” *Genome Research* 22 (4): 746–54. doi:10.1101/gr.125864.111.

McCormack, John E., Michael G. Harvey, Brant C. Faircloth, Nicholas G. Crawford, Travis C. Glenn, and Robb T. Brumfield. 2013a. “A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing.” *PLoS ONE* 8 (1): e54848. doi:10.1371/journal.pone.0054848.

———. 2013b. “Data from: A Phylogeny of Birds Based on over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing.” *Dryad Digital Repository*, January. http://hdl.handle.net/10255/dryad.45161.

Minin, Vladimir, Zaid Abdo, Paul Joyce, and Jack Sullivan. 2003. “Performance-Based Selection of Likelihood Models for Phylogeny Estimation.” *Systematic Biology* 52 (5): 674–83. doi:10.1080/10635150390235494.

Misof, Bernhard, Shanlin Liu, Karen Meusemann, Ralph S. Peters, Alexander Donath, Christoph Mayer, Paul B. Frandsen, et al. 2014. “Phylogenomics Resolves the Timing and Pattern of Insect Evolution.” *Science* 346 (6210): 763–67. doi:10.1126/science.1257570.

Nishihara, Hidenori, Norihiro Okada, and Masami Hasegawa. 2007. “Rooting the Eutherian Tree: The Power and Pitfalls of Phylogenomics.” *Genome Biology* 8 (9): R199. doi:10.1186/gb-2007-8-9-r199.

Nylander, Johan A. A., Fredrik Ronquist, John P. Huelsenbeck, and Joséluis Nieves-Aldrey. 2004. “Bayesian Phylogenetic Analysis of Combined Data.” *Systematic Biology* 53 (1): 47–67. doi:10.1080/10635150490264699.

Oaks, Jamie R. 2011a. “Data from: A Time-Calibrated Species Tree of Crocodylia Reveals a Recent Radiation of the True Crocodiles.” *Dryad Digital Repository*, June. http://hdl.handle.net/10255/dryad.33833.

———. 2011b. “A Time-Calibrated Species Tree of Crocodylia Reveals a Recent Radiation of the True Crocodiles.” *Evolution* 65 (11): 3285–97. doi:10.1111/j.1558-5646.2011.01373.x.

Ostrovsky, R., Y. Rabani, L.J. Schulman, and C. Swamy. 2006. “The Effectiveness of Lloyd-Type Methods for the K-Means Problem.” In *47th Annual IEEE Symposium on Foundations of Computer Science, 2006. FOCS ’06*, 165–76. doi:10.1109/FOCS.2006.75.

Pagel, Mark, and Andrew Meade. 2004. “A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data.” *Systematic Biology* 53 (4): 571–81. doi:10.1080/10635150490468675.

Paul Frandsen, and Christoph Mayer. n.d. *Fast-TIGER* (version v0.0.2). http://dx.doi.org/10.5281/zenodo.12914.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. *Scikit-Learn: Machine Learning in Python*. ArXiv e-print 1201.0490. http://arxiv.org/abs/1201.0490.

Phillips, Matthew J., Frédéric Delsuc, and David Penny. 2004. “Genome-Scale Phylogeny and the Detection of Systematic Biases.” *Molecular Biology and Evolution* 21 (7): 1455–58. doi:10.1093/molbev/msh137.

Posada, David, and Keith A. Crandall. 2001. “Selecting the Best-Fit Model of Nucleotide Substitution.” *Systematic Biology* 50 (4): 580–601. doi:10.1080/10635150118469.

Quang, Le Si, Olivier Gascuel, and Nicolas Lartillot. 2008. “Empirical Profile Mixture Models for Phylogenetic Reconstruction.” *Bioinformatics* 24 (20): 2317–23. doi:10.1093/bioinformatics/btn445.

Revell, Liam J., Luke J. Harmon, and Richard E. Glor. 2005. “Under-Parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies.” *Systematic Biology* 54 (6): 973–83. doi:10.1080/10635150500354647.

Robinson, D. F., and L. R. Foulds. 1981. “Comparison of Phylogenetic Trees.” *Mathematical Biosciences* 53 (1–2): 131–47. doi:10.1016/0025-5564(81)90043-2.

Rodríguez-Ezpeleta, Naiara, Henner Brinkmann, Béatrice Roure, Nicolas Lartillot, B. Franz Lang, and Hervé Philippe. 2007. “Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies.” *Systematic Biology* 56 (3): 389–99. doi:10.1080/10635150701397643.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012. “MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space.” *Systematic Biology* 61 (3): 539–42. doi:10.1093/sysbio/sys029.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model.” *The Annals of Statistics* 6 (2): 461–64. doi:10.1214/aos/1176344136.

Shapiro, Beth, Andrew Rambaut, and Alexei J. Drummond. 2006. “Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences.” *Molecular Biology and Evolution* 23 (1): 7–9. doi:10.1093/molbev/msj021.

Sharanowski, Barbara J., Ashley P. G. Dowling, and Michael J. Sharkey. 2011a. “Data from: Molecular Phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea) Based on Multiple Nuclear Genes and Implications for Classification.” *Dryad Digital Repository*, June. http://hdl.handle.net/10255/dryad.33714.

———. 2011b. “Molecular Phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea), Based on Multiple Nuclear Genes, and Implications for Classification.” *Systematic Entomology* 36 (3): 549–72. doi:10.1111/j.1365-3113.2011.00580.x.

Simon, Chris, Thomas R Buckley, Francesco Frati, James B Stewart, and Andrew T Beckenbach. 2006. “Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA.” *Annual Reviews in Ecology Evolution and Systematics* 37: 547–79.

Simon, Chris, Francesco Frati, Andrew Beckenbach, Bernie Crespi, Hong Liu, and Paul 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 (6): 651–701.

Simon, C, L Nigro, J Sullivan, K Holsinger, A Martin, A Grapputo, A Franke, and C McIntosh. 1996. “Large Differences in Substitutional Pattern and Evolutionary Rate of 12S Ribosomal RNA Genes.” *Molecular Biology and Evolution* 13 (7): 923–32.

Soubrier, Julien, Mike Steel, Michael S. Y. Lee, Clio Der Sarkissian, Stéphane Guindon, Simon Y. W. Ho, and Alan Cooper. 2012. “The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates.” *Molecular Biology and Evolution* 29 (11): 3345–58. doi:10.1093/molbev/mss140.

Springer, M S, H M Amrine, A Burk, and M J Stanhope. 1999. “Additional Support for Afrotheria and Paenungulata, the Performance of Mitochondrial versus Nuclear Genes, and the Impact of Data Partitions with Heterogeneous Base Composition.” *Systematic Biology* 48 (1): 65–75.

Stamatakis, Alexandros. 2006. “RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

———. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics*, January, btu033. doi:10.1093/bioinformatics/btu033.

Steel, Mike. 2005. “Should Phylogenetic Models Be Trying to ‘fit an Elephant’?.” *Trends in Genetics* 21 (6): 307–9. doi:10.1016/j.tig.2005.04.001.

Stergachis, Andrew B., Eric Haugen, Anthony Shafer, Wenqing Fu, Benjamin Vernot, Alex Reynolds, Anthony Raubitschek, et al. 2013. “Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution.” *Science* 342 (6164): 1367–72. doi:10.1126/science.1243490.

Sullivan, Jack, and Paul Joyce. 2005. “Model Selection in Phylogenetics.” *Annual Review of Ecology, Evolution, and Systematics* 36 (1): 445–66. doi:10.1146/annurev.ecolsys.36.102003.152633.

T. Grande, W. C. Borden. 2013. “Limits and Relationships of the Paracanthopterygii. A Molecular Framework for Evaluating Past Morphological Hypotheses.” *Mesozoic Fishes* 5: 385–418.

Wu, Chieh-Hsi, Marc A. Suchard, and Alexei J. Drummond. 2013. “Bayesian Selection of Nucleotide Substitution Models and Their Site Assignments.” *Molecular Biology and Evolution* 30 (3): 669–88. doi:10.1093/molbev/mss258.

Yang, Ziheng. 1994. “Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods.” *Journal of Molecular Evolution* 39 (3): 306–14. doi:10.1007/BF00160154.

———. 1996a. “Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data.” *Journal of Molecular Evolution* 42 (5): 587–96. doi:10.1007/BF02352289.

———. 1996b. “Among-Site Rate Variation and Its Impact on Phylogenetic Analyses.” *Trends in Ecology & Evolution* 11 (9): 367–72. doi:10.1016/0169-5347(96)10041-0.

Yang, Ziheng, and Bruce Rannala. 2012. “Molecular Phylogenetics: Principles and Practice.” *Nature Reviews Genetics* 13 (5): 303–14. doi:10.1038/nrg3186.

Zwickl, Derrick Joel. 2006. “Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion.” http://repositories.lib.utexas.edu/handle/2152/2666.

CHAPTER 3: PHYLOGENY OF PHRYGANEIDAE GENERA

Introduction

Phryganeidae biology and systematics

The giant case making caddisflies (Trichoptera: Phryganeidae) are some of the world's largest and most colorful caddisflies. Phryganeidae comprise roughly 80 described species in 15 genera (Morse 2009). Like all Trichoptera, their larvae are strictly aquatic. As part of the suborder Integripalpia, the larvae construct portable tube cases and can be found in cool streams and rivers, lakes, ephemeral pools, and even brackish environments (Holzenthal et al. 2007; Wiggins 1998). Phryganeidae are solely Holarctic, primarily occurring in northern latitudes (Holzenthal et al. 2007; Wiggins 1998). Some of the earliest caddisfly fossils belong to the Phryganeidae, dating back to the Lower Cretaceous (Sukatsheva 1968).

The history of taxonomic research on Phryganeidae has been rich, beginning with Linnaeus, who placed 15 caddisfly species in the genus *Phryganea* in his order Neuroptera (Wiggins 1998), three of which remain in the family Phryganeidae: *Phryganea grandis* (type species), *P. phalaenoides* (now *Semblis phalaenoides*), and *P. striata* (now *Oligotricha striata*) (Linnaeus 1758). Efforts to classify Phryganeidae continued through the work of Trichoptera experts such as Silfvenius (Silfvenius 1902; Silfvenius 1903; Silfvenius 1904), Martynov (Martynov 1924a; Martynov 1924b), Banks (Banks 1904; Banks 1914; Banks 1951), Schmid (Schmid 1962; Schmid 1965; Schmid 1968), and Wiggins

(Wiggins 1956; Wiggins 1960b; Wiggins 1960a; Wiggins 1972; Wiggins and Larson 1989) culminating in Wiggins' book written entirely about the family as we know it today (Wiggins 1998)—the only treatise of its kind published for a trichopteran family. In it, Wiggins described phryganeid life history, posited ideas about their historical biogeography, and proposed a hypothesis for their phylogeny based on pupal morphology. This phylogeny included what he considered strong evidence for some relationships, but the placement of a number of genera were more difficult to establish. Wiggins' phylogeny stands as the most recent phylogenetic hypothesis (Fig. 1).

Previous research on Phryganeidae suggests that the family consists of two subfamilies, Yphriinae comprised solely of the enigmatic species, *Yphria californica*, and Phryganeinae, which includes the other 14 genera (Wiggins 1962; Wiggins 1998). Phryganeid genera are relatively small in species diversity when compared with some of the other trichopteran genera. Four of the genera are monotypic and include *Beothukus*, *Fabria*, *Trichostegia*, and *Yphria*. The largest genus is Agrypnia, with 18 species (Morse 2009). In contrast, some of the largest genera in Trichoptera, e.g. *Chimarra* and *Rhyacophila*, contain over 700 and 400 species respectively (Morse 2009).

Some of the oldest caddisfly fossils are thought to belong to Phryganeidae. They consist of wing impressions from Mongolia dating back to the Lower

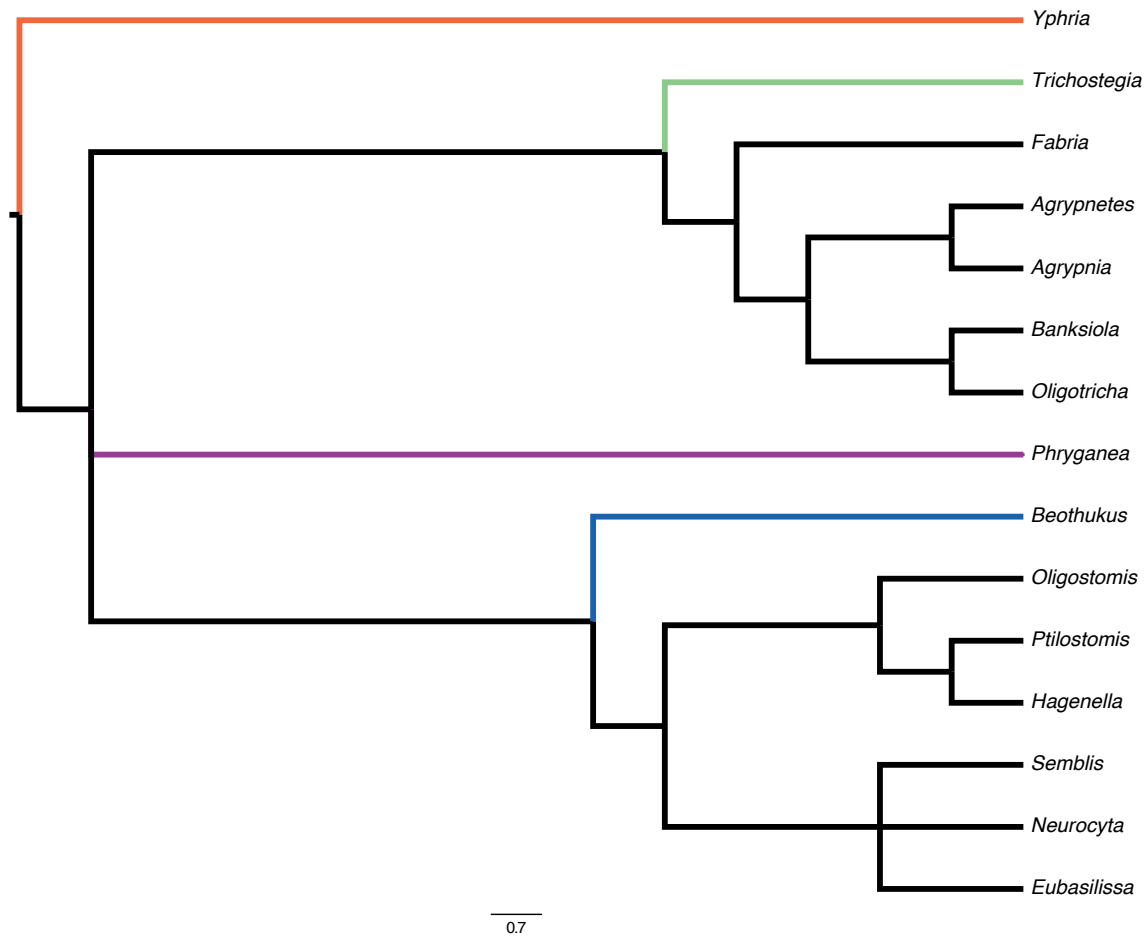

Figure 1. Phylogenetic hypothesis estimated by Wiggins (1998) using pupal morphology

Cretaceous (144-100 Ma) (Sukatsheva 1992) and larval case fossils with the characteristic spiral construction dating to the Upper Cretaceous (100-95 Ma) (Sukatsheva 1980). This indicates that the lineage may be at least that old and could have arisen following the rifting of Pangaea during the Jurassic, confining Phryganeidae to Laurasia. This is further supported by the absence of phryganeid fossils in the southern hemisphere (Wiggins 1998).

Three general morphological conditions exist for the larval cases within Phryganeidae: spiral cases, ring cases, and irregular cases (Fig. 2). The spiral and ring cases are considered to be the more derived states within the family since the genera with irregular cases, *Yphria* and *Trichostegia*, have been found to be basal with respect to the rest of the family (Wiggins 1998). Spiral and ringed case morphology is rare within Trichoptera (Hinchliffe and Palmer 2010; Wiggins 1998), and the remaining genera within Phryganeidae can be split into two groups according to their case construction behavior. However, some confusion exists as to the origin of the ring and spiral cases. Classification based on this morphological characteristic is not strongly supported due to the fact that *Phryganea* (which make spiral cases) and *Beothukus* (which make ring cases) share other morphological characteristics that are thought to be apomorphic, such as the compressed lateral lobes of the sub genital plate (Wiggins 1998). Wiggins hypothesized that the similarities in the morphology of the sub genital plate could be explained if ring cases were derived from spiral cases with *Phryganea* as sister to all ring cased genera. However, this

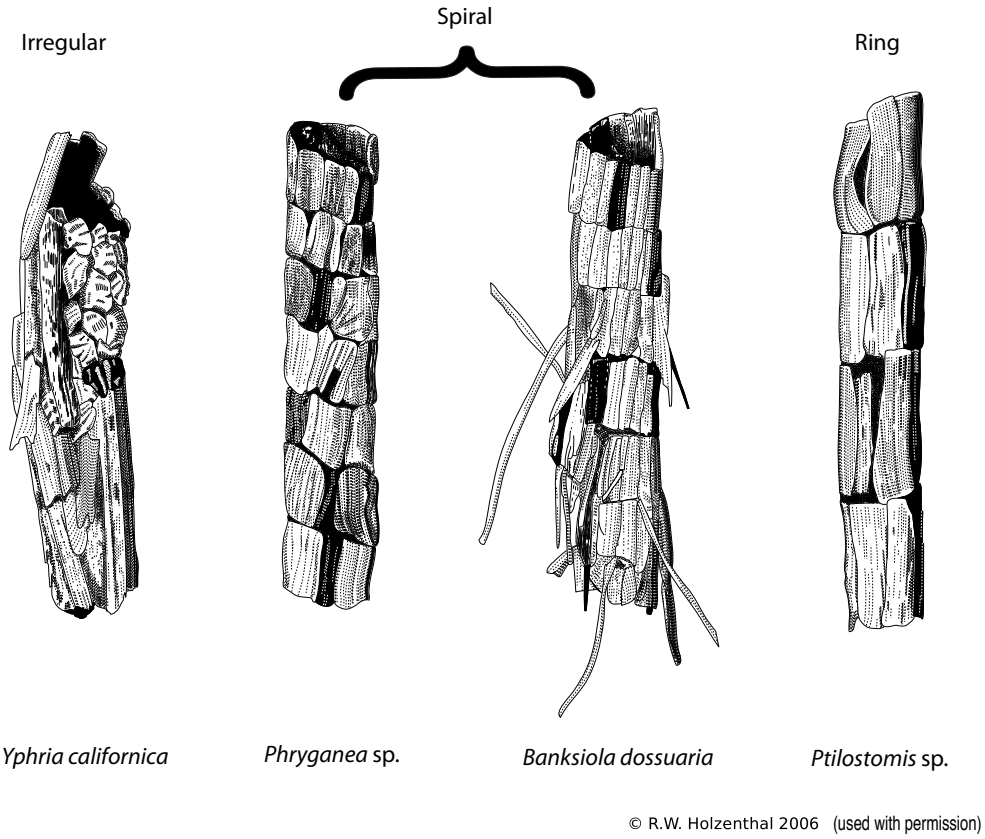

Figure 2. Illustrations of the three case morphologies for genera within Phryganeidae. Figure modified and reproduced from (Merritt, Cummins, and Berg 2008) with permission from the artist.

hypothesis conflicted with the morphology-derived parsimony tree he presented. While a close affinity between *Phryganea* and *Beothukus* seems plausible due to shared morphological characteristics, the disparity between their case morphology seems to suggest it is equally plausible that they share an evolutionary affinity with genera that possess the same case morphologies. Wiggins also expressed confusion over the placement of *Trichostegia*. Although *Trichostegia* lack a ring or spiral case, they share many other morphological characteristics with the genera within the subfamily Phryganeinae. To further complicate matters, *Yphria*, *Trichostegia*, and *Beothukus* are all monotypic, which can complicate the identification of shared morphological characteristics within the genera, rendering the classification of such groups difficult (Williams 2013). Thus, *Phryganea*, *Beothukus*, and *Trichostegia* remain unresolved in regards to their placement within Phryganeidae. We aim to test these conflicting hypotheses in the present study.

Targeted enrichment and high throughput sequencing

High throughput sequencing of short reads of DNA has revolutionized the generation of molecular data by massively decreasing the cost of producing data while simultaneously exponentially increasing the amount of molecular data that researchers can generate for their taxon of interest (Metzker 2010; Mardis 2013; van Dijk et al. 2014). Studies that have leveraged these technologies for phylogenetics have generally relied upon three main techniques: whole genome sequencing, transcriptome sequencing, and the selected enrichment of

preselected loci through the use of capture probes often referred to as “targeted enrichment”.

The use of RNAseq—the collection of total mRNA and subsequent construction and sequencing of cDNA libraries to generate transcriptomes—represents a powerful and comparatively inexpensive alternative to whole genome sequencing for phylogenetic data. Misof *et al.* found that when comparing across Arthropoda, the number of genes that can be recovered and identified as single copy orthologs using transcriptomes was similar to the number of single copy orthologs that could be found across genomes (Misof et al. 2014). Thus, the use of transcriptomes can potentially generate the same amount of phylogenetic data as genomes for a fraction of the cost. Several insect phylogenetic studies have taken advantage of this new technology with great success (Peters et al. 2014; Misof et al. 2014; Kawahara and Breinholt 2014). However, because of the ephemeral nature of mRNA, the generation of RNAseq data requires the collection and preservation of live material. Many labs have collections of tissues from thousands of species resulting from decades of collecting efforts and often containing rare or old material, all unsuitable for RNAseq. Sequencing genomes still requires a large amount of quality DNA and is cost prohibitive for dense taxon sampling, rendering it difficult for old and rare tissues and studies requiring the assessment of many taxa. For the present study, because of the lack of fresh material and the impossibility of generating genomes, we decided to generate data using a targeted enrichment strategy.

Targeted enrichment strategies were originally developed in order to sequence full human exomes. Known as “exome capture”, it has been used since 2009 in human health studies (Ng et al. 2009; Choi et al. 2009), allowing researchers to inexpensively sequence the protein coding genes in humans (~1% of total DNA content) (Perkel 2013). Because of a paucity of data on which to design capture probes outside of humans and due to nascent improvements in multiplexed sequencing of short read DNA, this technique has only recently been adopted for research on other organisms (Faircloth et al. 2012; Lemmon, Emme, and Lemmon 2012; McCormack et al. 2013; Crawford et al. 2012). These methods rely on decreasing the volume of DNA being sequenced by designing capture probes that hybridize to targeted sections of DNA. The sections not captured are then discarded and only the DNA of interest is sequenced. In order to ensure a successful capture with a high rate of hybridization, highly conserved areas in genomes across taxa must be targeted. This technique has been successfully applied to ultra-conserved elements (Faircloth et al. 2012), highly conserved areas of protein coding genes (Lemmon, Emme, and Lemmon 2012), and on more divergent loci aided by multiple hybridization cycles during the capture process (Li et al. 2013).

Here we use targeted enrichment techniques with probes designed based on conservative areas of protein coding genes from transcriptomes to estimate the phylogeny for Phryganeidae. We examine the utility and difficulties of designing

targeted enrichment probes from transcriptomes and discuss the challenges and opportunities that arise when estimating model-based phylogenies on large datasets. Finally we present a phylogeny of the genera within Phryganeidae.

Materials and Methods

Transcriptome generation and analysis

High throughput sequencing of targeted enriched data requires the design of capture probes for the enrichment of the target data. The design of proper capture probes requires a template from which to design them. Since at the time of this research, no Trichoptera genomes were available and in order to maintain compatibility with the transcriptomes that we were simultaneously collecting, we used transcriptomes to design our probes. The ideal templates for probe design are well-annotated genomes, which contain information about the full suite of genomic features, including the location of introns. Using genomes, researchers can design probes that will capture UCE's, intron elements flanked by conserved regions, and other non-protein coding features. We had no such luxury so we generated transcriptomes from which to design our capture probes. Since we wished to use this probe set on families outside of Phryganeidae and since a high quality capture relies on successful hybridization of target DNA to the probes, we carefully selected 15 taxa for transcriptome sequencing from throughout the order to widen the scope of the probe set and increase the probability of success.

The methods for transcriptome collection, extraction, library prep, sequencing, assembly, and orthology prediction followed the pipeline designed by the 1KITE project (Misof et al. 2014). Specimens were collected and macerated in RNAlater and stored in a refrigerator until sequencing in order to avoid excessive freeze/thaw cycles. The identification of live Trichoptera can be problematic since many of the features used to delineate species are only visible once the genitalia have been removed and treated properly (Blahnik, Holzenthal, and Prather 2007). To deal with this limitation, if possible, we used a single trichopteran specimen. Species identification was confirmed by comparison of mitochondrial COI sequences with identified species from the barcode of life database (Ratnasingham and Hebert 2007). When this was not possible (for small specimens due to a lack of material), we carefully sorted specimens into morphospecies while alive on ice and pooled individuals from the same morphospecies together.

After collection, total RNA extractions followed by cDNA library preparation and amplification was performed at BGI-Shenzhen. Following extraction and library prep, all samples were sequenced on an Illumina HiSeq 2000 with 150 bp paired end reads. The sequences were assembled using SOAPdenovo trans (Xie et al. 2014). Following the assembly of the transcripts, we determined orthologs using a core ortholog set for holometabolous insects mined from OrthoDB and best reciprocal hit blast in a modified version of hamstr (Ebersberger, Strauss, and Haeseler 2009; Misof et al. 2014).

Targeted enrichment

Since our transcriptomes mainly consisted of mRNA transcripts, we didn't have information about intron/exon splice sites. No Trichoptera genome had been sequenced at the time of this study and capture efficiency is compromised by poor hybridization for probes that span intron/exon splice sites. In an attempt to overcome this limitation, we used the genome annotations from the Giant Silkworm Moth (*Bombyx mori*) to map probable intron and exon boundaries onto the transcripts under the assumption that many of these intron sites would be shared between our Trichoptera specimens and *Bombyx mori*, a member of Lepidoptera, the sister order to Trichoptera. We then used a sliding window of 10 bp across each exon to estimate sequence conservation among the 15 taxa in our alignments. We used this information to select a set of exons that contained highly conserved areas flanked by more variable areas, which are useful for determining phylogenetic relationships. Our final list included 989 exons for which we designed probes within the Phryganeidae. Probes were manufactured by Agilent using the SureSelect protocol.

We selected taxa from each genus for which we had available tissues within Phryganeidae. We were unable to secure tissues for one genus, *Neurocyta*. Every other genus in the family was represented in our sample. DNA was extracted from each specimen using a G-Biosciences OmniPrep™ extraction kit. All samples were measured for double stranded DNA concentration using the

Life Technologies Qubit® system. We then estimated DNA fragment size by adding loading dye to 4 μL of DNA and visualizing the samples on a 1% agarose gel run for 1 hour with 120 volts.

For the library prep and targeted capture steps, we followed the protocol outlined by Meyer and Kircher (2010). In short, DNA was diluted to ~200-500ng/50μL per sample, then sonicated into ~500 bp fragments followed by blunt end repair. Illumina adapters were then ligated to the DNA fragments followed by index addition via PCR. Illumina indices allow for multiplexed sequencing by appending a unique index to the DNA fragments from each sample. To avoid misassignment of reads due to sequencing error, each index differed by at least two base pairs from every other sample index. We then added biotinylated capture probes to hybridize to the target DNA. Magnetic streptavidin beads, which bind to the biotinylated probes, were added to the solution and target regions were extracted using a magnet. Non-target DNA was discarded. Target DNA was then amplified, pooled, and sequenced on an Illumina HiSeq 2000 with 150 bp paired end reads.

Read assembly, orthology prediction, and alignment

Reads were assembled to the probe sequences using a custom assembly method described by Lemmon et al. (Lemmon, Emme, and Lemmon 2012). Using the reference probe sequences, reads were first aligned to the reference

probe sequences, then additional reads were locally assembled to extend past the probe regions.

When more than one ortholog was assembled to a reference probe within a single taxon, we selected the best ortholog by clustering contigs based on their pairwise distances. Using a pairwise distance matrix, a neighbor-joining tree was constructed to determine orthologous clusters. In the cases that orthology was difficult to assign because two orthologs were equally probable, we took the conservative measure of removing both, leaving the taxon without data for that particular locus. We also removed clusters that were missing more than 50% of the species from all downstream analyses.

Following assembly and orthology sorting, we constructed single locus multiple sequence alignments using MAFFT 7.2 (Katoh and Standley 2013). Since loci often contained sequences that extended past the exon boundaries and into the intron, the locus alignments often contained large stretches of data that were difficult to align. Since misaligned data is similar to random data, we used Aliscore (Misof and Misof 2009), which uses a Monte Carlo resampling along a sliding window to identify sites in the alignment that are found to be indistinguishable from random data. We used a 5 bp sliding window and the "-N" option that treats gaps as missing data. Alternatively, gaps can be treated as a 5th character, which is sometimes applied for alignments that consist of rRNA where insertions and deletions are common and informative. Since all loci that

we targeted in this study were protein-coding genes and the introns were deemed to difficult to align, we treated gaps as missing data. Following the Aliscore step, we removed putative random sections with the Alicut 2.3 utility (Kueck 2013). Aliscore has been shown to successfully remove sections of the alignment that represent misalignments or random data, but it does not remove taxa that may contain data from a paralog or is otherwise misaligned, because Aliscore uses permutation to compare the data to randomized nucleotides of similar properties, and a single taxon or two would not be sufficient to "randomize" the otherwise informative data from other taxa. To check for taxa that were obviously misaligned or contained obvious paralogs, we examined each locus alignment, colored by translated amino acids, by eye in Geneious v. 7 (Kearse et al. 2012) and removed problematic taxa. We then created a super-matrix alignment by concatenating all loci using FASconCAT (Kück and Meusemann 2010).

Phylogenetic model selection and data exclusion

To select phylogenetic models of molecular evolution, we employed two different approaches. In the first, we selected models and a partitioning scheme using the relaxed clustering algorithm in PartitionFinder v. 1.1.0 (Lanfear et al. 2012; Lanfear et al. 2014). The relaxed clustering algorithm begins with a large number of user-defined data blocks that are then evaluated to determine key model parameter values (relative evolutionary rate, transition/transversion ratios, and nucleotide composition). Data blocks that evolve under similar models are

then iteratively lumped together and evaluated for improvements in model fit using AICc. When the lumping results in an improved model fit, it is accepted. This process reduces the number of overall partitions and parameters reducing the risk of overparameterization. The time required for an algorithm run increases rapidly with an increasing number of predefined data blocks and recent research has found that predefining data blocks based on locus are likely to give a similar result as predefining data blocks based on locus and codon position (Kainer and Lanfear 2015). Given these findings and the additional, potentially unnecessary computational burden, we specified each locus as an individual data block. In our second approach, we selected the partitioning scheme using the iterative *k*-means algorithm that has been implemented into a development version of PartitionFinder (Frandsen et al. 2015). Iterative *k*-means does not require the *a priori* definition of data blocks; rather it estimates an optimal partitioning scheme directly from the data using rates estimated from individual sites.

The most commonly used models in phylogenetics consist of continuous time Markov substitution models. The most generalized of these is the general time reversible (GTR) model (Tavaré 1986). GTR allows for different rates of transitions from each nucleotide state to every other. It also allows for unequal base frequencies. The only DNA substitution models currently available in RAxML are GTR models with either the discrete GAMMA mixture correction for among site rate variation (GTR+G) (Yang 1994; Yang 1996) or discrete GAMMA

plus an estimate of the proportion of invariable sites (GTR+I+G). GTR makes several important assumptions: time-reversibility, stationarity of processes, compositional homogeneity, and site-to-site independence. The assumptions of stationarity, reversibility, and homogeneity can be assessed using matched-pairs tests of symmetry (Bowker 1948; Stuart 1955; Ababneh et al. 2006; Misof et al. 2014). These tests must be performed on multiple sites and have been used previously on individual loci, codon positions, or contiguous blocks of data using a sliding window. Motivated by the observation that individual sites may violate model assumptions when the surrounding sites do not, we took an alternative approach and performed the tests on the model subsets estimated during the iterative *k*-means selection of the partitioning scheme, which contain sites that have evolved under similar models.

Phylogenetic tree estimation

We generated three maximum likelihood (ML) trees to evaluate congruence among the different methods: (1) partitioned by locus following relaxed clustering partitioning selection with the GTR+G model applied to each subset, referred to as Tree 1, (2) partitioned by iterative *k*-means with the GTR+G model applied to each subset, referred to as Tree 2, and (3) partitioned by iterative *k*-means with the exclusion of subsets found to be in violation of any of the three model assumptions that we tested, referred to as Tree 3. The GTR+G model was applied to the resulting subsets. ML trees were estimated in RAxML 8 (Stamatakis 2014) with 5 independent thorough searches of each dataset. The

Table 1. Taxa used in this study

Family	Species	Author/Date
Brachycentridae	*Brachycentrus americanus*	Banks 1899
Lepidostomatidae	*Lepidostoma hirtum*	Fabricius 1775
Goeridae	*Goeracea oregona*	Denning 1968
Uenoidae	*Farula praelonga*	Wiggins & Erman 1987
Apataniidae	*Apataniana hellenica*	Malicky 1987
Limnephilidae	*Annitella thuringica*	Ulmer 1909
Limnephilidae	Limnephilidae sp.	N/A
Phryganeidae	*Yphria californica*	Banks 1907
Phryganeidae	*Trichostegia minor*	Curtis 1834
Phryganeidae	*Beothukus complicatus*	Banks 1924
Phryganeidae	*Phryganea bipunctata*	Retzius 1783
Phryganeidae	*Phryganea grandis**	Linnaeus 1758
Phryganeidae	*Hagenella apicalis*	Matsumura 1904
Phryganeidae	*Oligostomis ocelligera*	Walker 1858
Phryganeidae	*Ptilostomis semifasciata**	Leach 1815
Phryganeidae	*Eubasilissa maclachlani*	White 1862
Phryganeidae	*Semblis melaleuca*	McLachlan 1871
Phryganeidae	*Semblis phalaenoides*	Linnaeus 1758
Phryganeidae	*Banksiola crotchi*	Banks 1944
Phryganeidae	*Oligotricha striata*	Linnaeus 1758
Phryganeidae	*Agrypnia straminea*	Hagen 1873
Phryganeidae	*Agrypnia obsoleta*	Hagen 1864
Phryganeidae	*Agrypnia czerskyi*[#]	Martynov 1824

*-indicates sample sequences were derived from transcriptomes, #-two specimens were included from this species

tree with the best likelihood was chosen as the ML tree. We used the RAxML rapid bootstrap algorithm and generated bootstrap trees from 100 bootstrap alignments for each data set. The bootstraps were then projected on the final ML tree to reflect support.

Results

Targeted enrichment results

Of the 989 loci that we targeted within Phryganeidae, we recovered 892 loci that contained 13 or more taxa (out of 26 targeted) resulting in a roughly 90% recovery rate. For three samples, two taken from a single *Fabria inornata* specimen collected in the 1960s and one from an undetermined *Banksiola* species, we recovered only a handful of loci, which consisted almost entirely of non-trichopteran DNA. These taxa were removed, resulting in another missing genus (*Fabria*) from our study. Another taxon, *Oligotricha spicata* was determined to largely consist of contaminants and was also removed. We added the exon sequences from the transcriptomes of two species, *Ptilostomis semifasciata* and *Phryganea grandis*, from which we designed the target probes. This brought the total number of taxa to 24, including seven outgroup taxa (Table 1). In the final locus alignments, the average number of taxa recovered per locus was 22 of 24 (~91.6%). The average locus length was 650 bp (Fig. 3). Although the capture failed on one of our oldest specimens (*Fabria inornata*), we recovered DNA from several pinned museum specimens, including many that were over a decade old.

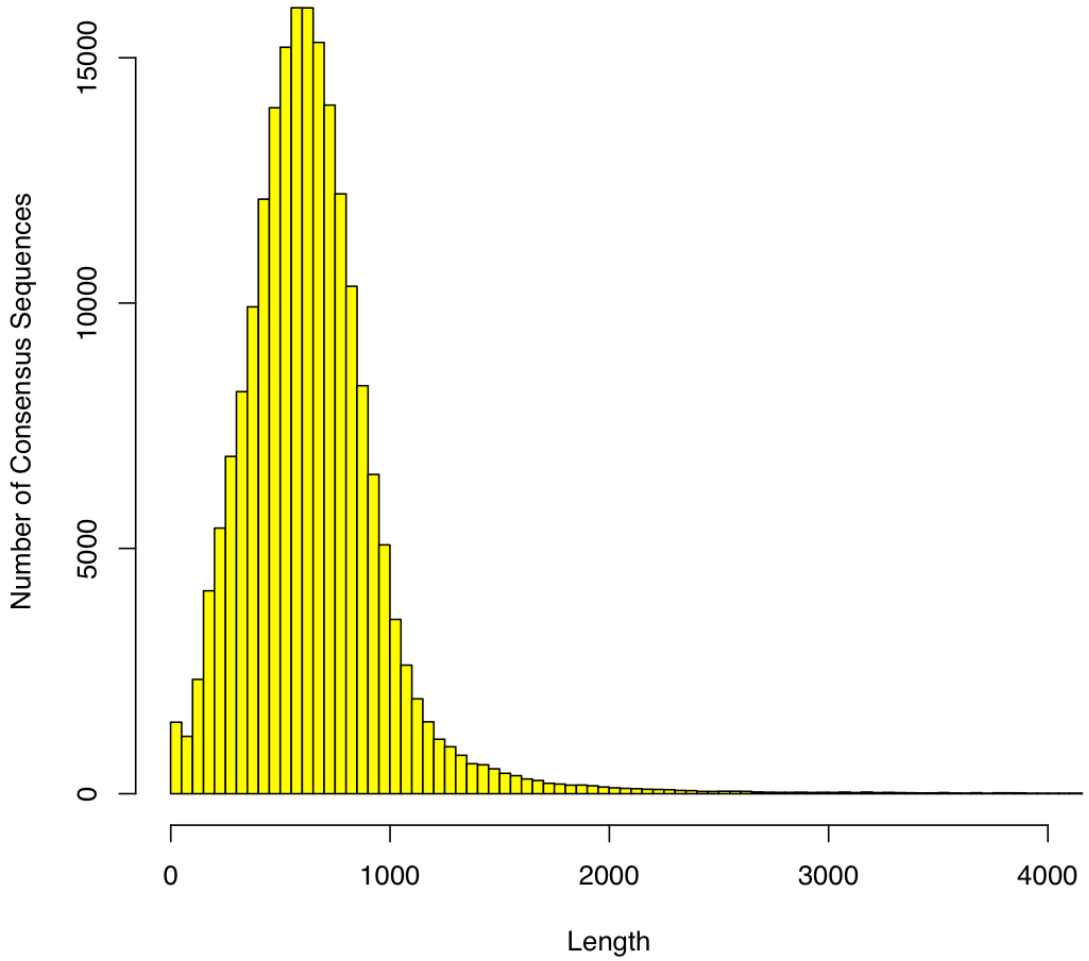

Figure 3. Histogram portraying the average locus length generated from the targeted enrichment sequencing.

After alignment masking using Aliscore (Misof and Misof 2009), some loci were found to include no sites that were distinguishable from random data. As a result, we removed 53 loci trimming the total number of loci to 839. Following concatenation, the final alignment included 24 taxa and 272,854 nucleotide positions.

We generated three trees to compare model selection and data exclusion approaches. Two of these trees (Tree 1 and Tree 2) were estimated from the dataset described above each with a different partitioned model. Tree 1 was partitioned by locus then optimized with the PartitionFinder relaxed clustering algorithm. Tree 2 was partitioned using the PartitionFinder iterative *k*-means algorithm. The AICc score for the two partitioned models was 3,861,597 for the relaxed clustering partitioned model and 3,381,776 for the iterative *k*-means partitioned model (lower AICc scores indicate a better fit of the data to the model). AICc scores are based upon the maximum likelihood units and differences as small as 10 AICc units are considered enough to represent a large difference in the relative fit of each model. Therefore, the difference in AICc between these two datasets of 517,614 AICc units is a drastic improvement in the fit of the model to the data. The relaxed clustering algorithm found 333 subsets while the iterative *k*-means algorithm found 61. Using matched pairs tests (Ababneh et al. 2006; Stuart 1955; Bowker 1948), we found 11 subsets selected by the *k*-means algorithm to be in violation of model assumptions. We removed these from our third dataset resulting in a trimmed alignment length of

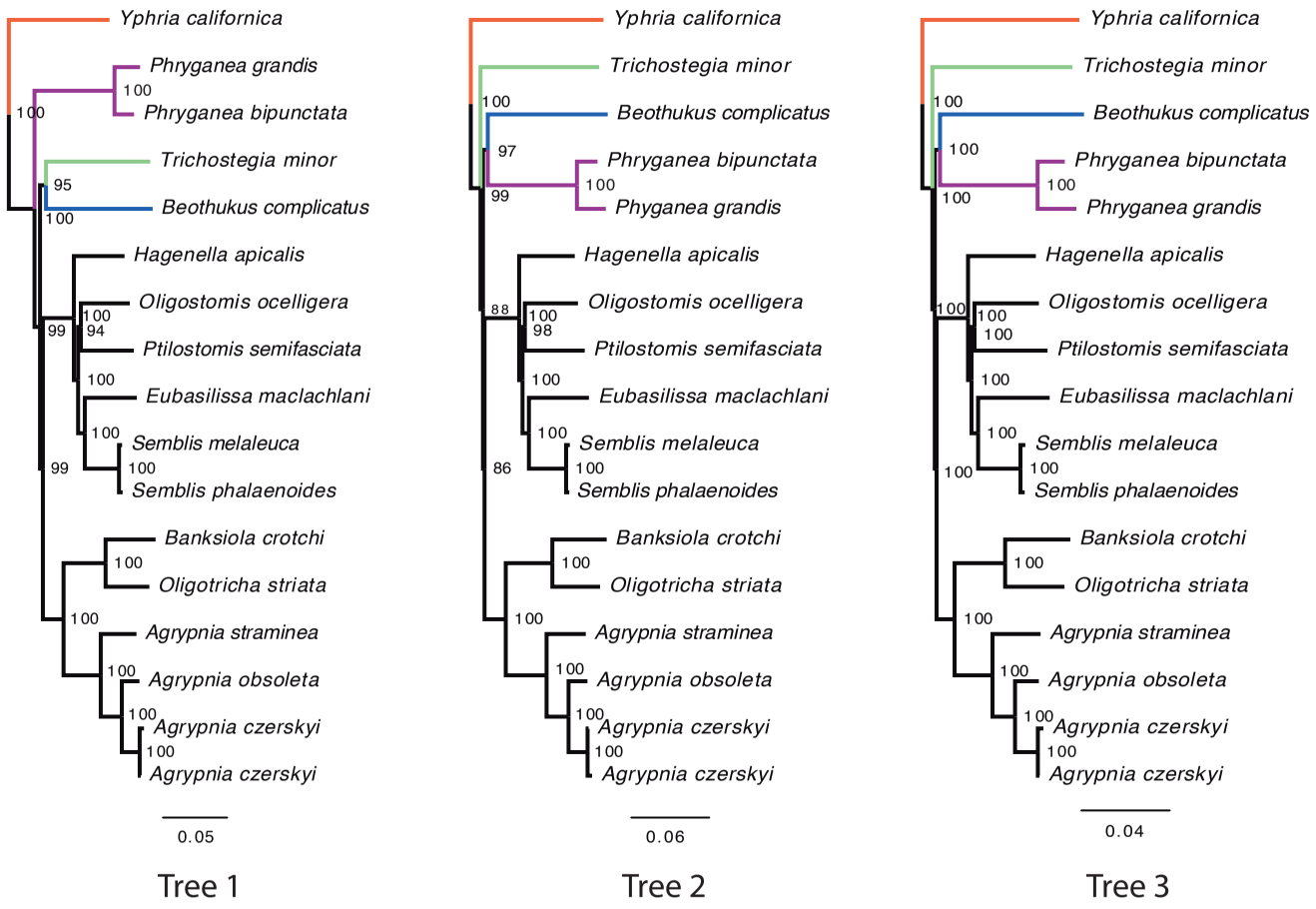

Figure 4. Maximum likelihood trees for the genera within Phryganeidae. Tree 1 was generated using the model selected using locus and the relaxed clustering algorithm, tree 2 was generated using the model selected via the iterative *k*-means algorithm, and tree 3 was generated using the model selected using the iterative *k*-means algorithm with model violating subsets removed. Outgroups were supported as monophyletic with 100 percent bootstraps for each topology and are not shown for clarity. The colored taxa represent "taxa of interest" given their uncertain placement in earlier phylogenies as well as their incongruent placement in our analyses.

254,070 bp (a difference of 18,784 bp). We generated Tree 3 from this trimmed data set. Due to the difference in alignment length, the model applied to this alignment cannot be compared to the others using information theoretic metrics like AICc.

The branching patterns of Tree 2 and Tree 3 (both partitioned using the iterative *k*-means algorithm) were congruent, differing mainly in support (Fig. 4). Tree 3 had the highest support with 100% bootstrap on every node. Tree 1 contained many nodes that were congruent with trees 2 and 3, but some important nodes were incongruent with the *k*-means trees (Fig. 4). All three methods recovered *Yphria* as the sister to the rest of Phryganeidae. In trees 2 and 3 *Trichostegia* is recovered as the next branch followed by a clade that includes the two *Phryganea* species as sister to *Beothukus*. Tree 1 differed at these nodes and included the two *Phryganea* species as the next bifurcation following *Yphria* with a clade containing *Beothukus* + *Trichostegia* recovered as sister to the rest of the genera in the family. All three methods recovered identical relationships for the remaining phryganeid genera (Fig. 4). Thus, the choice of partitioning scheme influenced the results.

Discussion

Targeted enrichment from transcriptomes

Here, we presented a method for generating a large amount of data at a relatively low cost for a group of organisms for which no reference genome

exists. We showed that by generating transcriptomes and by targeting portions of the transcripts that show a high level of conservation, it is possible to successfully capture loci over wide divergences. Ideally, the technology involved with generating genomes will continue to improve in quality and decrease in price rendering the sequencing of whole genomes tractable for phylogenetic studies involving many taxa. However, despite the impressive progress that has been made over the past decade with the advent of high throughput sequencing, whole genome sequencing is still infeasible for most studies. In addition, the computational burden associated with the generation, storage, and analysis of whole genomes remains high, and will likely always be significant. We found targeted enrichment to be an ideal, low-cost alternative to some of the other methods that leverage the power of high throughput sequencing, including low coverage genome sequencing and RNAseq.

We also showed that targeted enrichment could be successfully applied to sequence decade-old museum specimens. Many collections contain specimens that are rare and difficult to collect. Targeted enrichment provides an especially good opportunity to generate large amounts of data from these types of specimens. This is due to the serendipitous feature that the high throughput sequencing technology used relies on short fragments of DNA. Since degraded DNA is simply DNA that has already been fragmented, these technologies are well suited for old DNA.

Phylogeny of Phryganeidae

We generated three phylogenetic trees from three different treatments of the alignment. The two trees generated from the datasets for which models were selected using the iterative *k*-means algorithm had fully congruent topologies, differing only in the level of bootstrap support of a handful of nodes (Fig. 4). The support was highest on the tree generated from the alignment with model violating subsets removed (tree 3, Fig. 4). Interestingly, the tree generated from the data modeled with the relaxed clustering partitioning scheme also had high bootstrap support, but for a conflicting branching pattern (Fig. 4).

Phryganeidae was recovered as monophyletic in each analysis. All three trees also recovered *Yphria* as sister to the rest of Phryganeidae, confirming previous research (Wiggins 1962; Wiggins 1998) and providing further support for it's placement in its own subfamily, Yphriinae, and distinct from other Phryganeidae. The next two branches in the tree are incongruent between the topologies estimated using partitioning schemes generated with the iterative *k*-means and relaxed clustering algorithms. The taxa that are placed differently between the topologies are *Phryganea* and the monotypic genera *Beothukus* and *Trichostegia* with tree 1 recovering *Beothukus* as sister to *Trichostegia* and trees 2 and 3 recovering *Phryganea* as sister to *Trichostegia*. The analysis Wiggins conducted was also inconclusive regarding the placement of these taxa (Wiggins 1998). For example, he stated that there were some similarities that would suggest a sister relationship between *Phryganea* and *Beothukus*, such as

the compressed lateral lobes of the sub genital plate, which were determined to be an apomorphic condition. However, Wiggins also found strong support for *Beothukus* as sister to the rest of the ring case makers. None of our trees support the affinity of *Beothukus* to the ring case makers. Rather, we recover *Beothukus* as either sister to *Phryganea* or *Trichostegia*.

Early Phryganeidae evolution falls into the category of classically difficult phylogenetic problems. Early in the evolution of the family, there was a rapid diversification, indicated by the extremely short internodes in the base of the tree. This was followed by the apparent extinction of much of the stem lineage with few relictual, monotypic taxa remaining. This is similar to other phylogenetic problems in Insecta such as the proliferation and subsequent extinction of Polyneoptera (Whitfield and Kjer 2008). As such it is perhaps unsurprising, but still disconcerting, that variations in the selection of the model would generate two incongruent, highly supported topologies. In large data sets such as the one used for this study, it is important to keep in mind that the support values generated by the non-parametric bootstrap are not the probability of a node being true, rather they are a measure of confidence in the robustness of the tree to perturbations in the data matrix due to random resampling. It is impossible to determine with certainty which hypothesis represents the “true tree”, however, we can examine other lines of evidence that may lend support to one hypothesis over the other. Given this stipulation, there are two reasons that we are more confident in the tree estimated from the *k*-means partitioning schemes: (1) the fit

of the model to the data for the *k*-means selected model (AICc of 3,381,776) was far superior to the fit of the model selected using the relaxed clustering algorithm (AICc of 3,861,597), indicating that the model is capturing more information about the evolutionary process, and (2) morphological evidence from previous research seems to better support a relationship between *Beothukus* and *Phryganea* than that between *Beothukus* and *Trichostegia*.

The large improvement in the AICc score for the substitution model selected using iterative *k*-means when compared with the model selected using relaxed clustering suggests that for targeted enrichment datasets like these, using an automatic phenomenological model selection technique such as iterative *k*-means is likely to improve the information in the model as opposed to more commonly used methods such as straightforward mechanistic modeling by locus. We have also shown that the iterative *k*-means method can be used to successfully identify groups of sites that are in violation of model assumptions. In this dataset, the removal of these sites resulted in improved support via the non-parametric bootstrap (Felsenstein 1985) suggesting that the resulting data matrix was less biased by model misspecification and more robust to the bootstrap resampling with replacement process. For example, we would expect that if false nodes are supported and misleading data is excluded support values would decrease. Conversely, if nodes are correct and misleading data is eliminated, bootstrap support should improve.

The method that we used for identifying sites that violate model assumptions has potential for widespread use in phylogenetic analyses. Previous studies that have used matched pairs tests to remove sites have all relied on evaluating groups of sites based on locus or codon boundaries, or on evaluating contiguous groups of sites with variable sliding window sizes (Jayaswal et al. 2011; Ababneh et al. 2006; Misof et al. 2014). Using these methods to search for sites that violate model assumptions generalizes over a group of sites that could have evolved under very different patterns. In some cases, both 1st and 3rd codon positions failed matched pairs tests, resulting in 2/3 of the total data being removed (Misof et al. 2014), which at best resulted in too many sites being removed and at worst could bias the analyses (Lartillot and Philippe 2004; Kumar et al. 2012; Pagel and Meade 2004). In our approach, we applied the matched pairs tests to subsets of similar sites that were clustered together during the iterative *k*-means partitioning. Since the clustering algorithm treats sites individually (rather than on somewhat arbitrary sliding windows or locus/codon position boundaries), the subsets that were found to be in violation of the model assumptions are more likely to consist primarily of similar sites that are all in violation of model assumptions. By the same token, the iterative *k*-means approach is likely to group sites that evolve under similar conditions, which could result in more accurate modeling and a better fulfillment of the assumptions of homogeneity and stationarity. Applying these tests for data exclusion to subsets of similar sites could result in fewer sites being excluded

from the analysis due to violations by neighboring sites, or sites from the same biological group.

Conclusions and directions for future research

We find that targeted enrichment is a tractable option for the generation of large molecular datasets for non-model organisms. We also find strong support for the relationships among genera in the caddisfly family Phryganeidae. Although conflicting trees were estimated from different modeling strategies, we place more confidence on the tree estimated from the iterative *k*-means partitioning. We find that differences in modeling can generate strongly supported, conflicting topologies and caution researchers to explore many methods of modeling when estimating phylogenetic trees. A further exploration of the role of bootstraps in analyses of large datasets is an important area of further research. We also find that clustering together similar sites can help identify groups of sites in violation of model assumptions. We show that the removal of such sites can result in stronger support for a given topology.

References

Ababneh, Faisal, Lars S Jermiin, Chunsheng Ma, and John Robinson. 2006. "Matched-Pairs Tests of Homogeneity with Applications to Homologous Nucleotide Sequences." *Bioinformatics (Oxford, England)* 22 (10): 1225–31. doi:10.1093/bioinformatics/btl064.

Banks, Nathan. 1904. "A List of Neuropteroid Insects, Exclusive of Odonata, from the Vicinity of Washington, D. C." *Proceedings of the Entomological Society of Washington* 6: 201–17.

———. 1914. "American Trichoptera—Notes and Descriptions." *The Canadian Entomologist* 46 (07): 252–58. doi:10.4039/Ent46252-7.

———. 1951. “Notes on Some New England Phryganeidae (Trichoptera).” *Psyche: A Journal of Entomology* 58 (1): 20–23. doi:10.1155/1951/24581.

Blahnik, Roger J., Ralph W. Holzenthal, and Aysha Prather. 2007. “The Lactic Acid Method for Clearing Trichoptera Genitalia.” *Proceedings of the XIIth Symposium on Trichoptera*, The Caddis Press, , 9–14.

Bowker, A. H. 1948. “A Test for Symmetry in Contingency Tables.” *Journal of the American Statistical Association* 43 (244): 572–74.

Choi, Murim, Ute I. Scholl, Weizhen Ji, Tiewen Liu, Irina R. Tikhonova, Paul Zumbo, Ahmet Nayir, et al. 2009. “Genetic Diagnosis by Whole Exome Capture and Massively Parallel DNA Sequencing.” *Proceedings of the National Academy of Sciences* 106 (45): 19096–101. doi:10.1073/pnas.0910672106.

Crawford, Nicholas G., Brant C. Faircloth, John E. McCormack, Robb T. Brumfield, Kevin Winker, and Travis C. Glenn. 2012. “More than 1000 Ultraconserved Elements Provide Evidence That Turtles Are the Sister Group of Archosaurs.” *Biology Letters* 8 (5): 783–86. doi:10.1098/rsbl.2012.0331.

Ebersberger, Ingo, Sascha Strauss, and Arndt von Haeseler. 2009. “HaMStR: Profile Hidden Markov Model Based Search for Orthologs in ESTs.” *BMC Evolutionary Biology* 9 (1): 157. doi:10.1186/1471-2148-9-157.

Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales.” *Systematic Biology* 61 (5): 717–26. doi:10.1093/sysbio/sys004.

Felsenstein, Joseph. 1985. “Confidence Limits on Phylogenies: An Approach Using the Bootstrap.” *Evolution* 39: 783–91. doi:10.2307/2408678.

Frandsen, Paul B., Brett Calcott, Christoph Mayer, and Robert Lanfear. 2015. “Automatic Selection of Partitioning Schemes for Phylogenetic Analyses Using Iterative K-Means Clustering of Site Rates.” *BMC Evolutionary Biology* 15 (1): 13. doi:10.1186/s12862-015-0283-7.

Hinchliffe, Robert, and A. R. Palmer. 2010. “Curious Chiral Cases of Caddisfly Larvae: Handed Behavior, Asymmetric Forms, Evolutionary History.” *Integrative and Comparative Biology* 50 (4): 606–18. doi:10.1093/icb/icq069.

Holzenthal, Ralph W., Roger J. Blahnik, Aysha L. Prather, and Karl M. Kjer. 2007. “Order Trichoptera Kirby, 1813 (Insecta), Caddisflies.” *Zootaxa*, no. 1668 (December): 639–98.

Jayaswal, Vivek, Faisal Ababneh, Lars S Jermiin, and John Robinson. 2011. “Reducing Model Complexity of the General Markov Model of Evolution.” *Molecular Biology and Evolution* 28 (11): 3045–59. doi:10.1093/molbev/msr128.

Kainer, David, and Robert Lanfear. 2015. “The Effects of Partitioning on Phylogenetic Inference.” *Molecular Biology and Evolution*, February, msv026. doi:10.1093/molbev/msv026.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. doi:10.1093/molbev/mst010.

Kawahara, Akito Y., and Jesse W. Breinholt. 2014. “Phylogenomics Provides Strong Evidence for Relationships of Butterflies and Moths.” *Proceedings. Biological Sciences / The Royal Society* 281 (1788): 20140970. doi:10.1098/rspb.2014.0970.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. “Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data.” *Bioinformatics (Oxford, England)* 28 (12): 1647–49. doi:10.1093/bioinformatics/bts199.

Kück, Patrick, and Karen Meusemann. 2010. “FASconCAT: Convenient Handling of Data Matrices.” *Molecular Phylogenetics and Evolution* 56 (3): 1115–18. doi:10.1016/j.ympev.2010.04.024.

Kueck, Patrick. 2013. *Alicut* (version 2.3). https://www.zfmk.de/en/research/research-centres-and-groups/utilities.

Kumar, Sudhir, Alan J. Filipski, Fabia U. Battistuzzi, Sergei L. Kosakovsky Pond, and Koichiro Tamura. 2012. “Statistics and Truth in Phylogenomics.” *Molecular Biology and Evolution* 29 (2): 457–72. doi:10.1093/molbev/msr202.

Lanfear, Robert, Brett Calcott, Simon Y. W. Ho, and Stephane Guindon. 2012. “PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses.” *Molecular Biology and Evolution*, January, mss020. doi:10.1093/molbev/mss020.

Lanfear, Robert, Brett Calcott, David Kainer, Christoph Mayer, and Alexandros Stamatakis. 2014. “Selecting Optimal Partitioning Schemes for Phylogenomic Datasets.” *BMC Evolutionary Biology* 14 (1): 82. doi:10.1186/1471-2148-14-82.

Lartillot, Nicolas, and Hervé Philippe. 2004. “A Bayesian Mixture Model for across-Site Heterogeneities in the Amino-Acid Replacement Process.” *Molecular Biology and Evolution* 21 (6): 1095–1109. doi:10.1093/molbev/msh112.

Lemmon, Alan R., Sandra A. Emme, and Emily Moriarty Lemmon. 2012. “Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics.” *Systematic Biology* 61 (5): 727–44. doi:10.1093/sysbio/sys049.

Li, Chenhong, Michael Hofreiter, Nicolas Straube, Shannon Corrigan, and Gavin J. P. Naylor. 2013. “Capturing Protein-Coding Genes across Highly Divergent Species.” *BioTechniques* 54 (6): 321–26. doi:10.2144/000114039.

Linnaeus, Carl. 1758. *Systema Naturae per Regna Tira Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis,*

Synonymis, Locis. 10th ed. Vol. 1, Animalia. Stockholmiae: Laurentii Salvii.

Mardis, Elaine R. 2013. “Next-Generation Sequencing Platforms.” *Annual Review of Analytical Chemistry* 6 (1): 287–303. doi:10.1146/annurev-anchem-062012-092628.

Martynov, A.V. 1924a. “Preliminary Revision of the Family Phryganeidae, Its Classification and Evolution.” *Annals and Magazine of Natural History*, 9, 14: 209–24.

———. 1924b. “Sur la classification et l’évolution de la fam. Phryganeidae (Trichoptera).” *Comptes Rendes de l Académie des Sciences de Russie*, 77–80.

McCormack, John E., Michael G. Harvey, Brant C. Faircloth, Nicholas G. Crawford, Travis C. Glenn, and Robb T. Brumfield. 2013. “A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing.” *PLoS ONE* 8 (1): e54848. doi:10.1371/journal.pone.0054848.

Merritt, R. W., K. W. Cummins, and M. B. Berg, eds. 2008. *AN INTRODUCTION TO THE AQUATIC INSECTS OF NORTH AMERICA*. 4 edition. Dubuque, Iowa: Kendall Hunt Publishing.

Metzker, Michael L. 2010. “Sequencing Technologies - the next Generation.” *Nature Reviews. Genetics* 11 (1): 31–46. doi:10.1038/nrg2626.

Meyer, Matthias, and Martin Kircher. 2010. “Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing.” *Cold Spring Harbor Protocols* 2010 (6): pdb.prot5448. doi:10.1101/pdb.prot5448.

Misof, Bernhard, Shanlin Liu, Karen Meusemann, Ralph S. Peters, Alexander Donath, Christoph Mayer, Paul B. Frandsen, et al. 2014. “Phylogenomics Resolves the Timing and Pattern of Insect Evolution.” *Science* 346 (6210): 763–67. doi:10.1126/science.1257570.

Misof, Bernhard, and Katharina Misof. 2009. “A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments: A More Objective Means of Data Exclusion.” *Systematic Biology* 58 (1): 21–34. doi:10.1093/sysbio/syp006.

Morse, John C. 2009. “The Trichoptera World Checklist.” *Zoosymposia*. http://works.bepress.com/john_morse/24.

Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, et al. 2009. “Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes.” *Nature* 461 (7261): 272–76. doi:10.1038/nature08250.

Pagel, Mark, and Andrew Meade. 2004. “A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data.” *Systematic Biology* 53 (4): 571–81. doi:10.1080/10635150490468675.

Perkel, Jeffrey M. 2013. “Exome Sequencing: Toward an Interpretable Genome.” AAAS. http://www.sciencemag.org/site/products/lst_20131011.pdf.

Peters, Ralph S., Karen Meusemann, Malte Petersen, Christoph Mayer, Jeanne Wilbrandt, Tanja Ziesmann, Alexander Donath, et al. 2014. “The Evolutionary History of Holometabolous Insects Inferred from Transcriptome-Based Phylogeny and Comprehensive Morphological Data.” *BMC Evolutionary Biology* 14 (1): 52. doi:10.1186/1471-2148-14-52.

Ratnasingham, Sujeevan, and Paul D N Hebert. 2007. “Bold: The Barcode of Life Data System (http://www.barcodinglife.org).” *Molecular Ecology Notes* 7 (3): 355–64. doi:10.1111/j.1471-8286.2007.01678.x.

Schmid, F. 1962. “Le Genre Eubasilissa Mart. En Inde.” *Bulletin de La Société Vaudoise Des Sciences Naturelles* 68 (309): 153–68.

———. 1965. “Encore Une Eubasilissa Himalayenne.” *Canadian Entomologist* 97 (1): 108–9.

———. 1968. “Les Genres Neurocyta Navas et Phryganopsyche Wiggins En Inde.” *Naturaliste Canadien* 95: 723–26.

Silfvenius, A.J. 1902. “Über Die Matamorphose Einiger Phryganeiden Und Limnophiliden.” *Acta Societas pro Fauna et Flora Fennica* 21 (4): 1–102.

———. 1903. “Über Die Matamorphose Einiger Phryganeiden Und Limnophiliden II.” *Acta Societas pro Fauna et Flora Fennica* 25 (4): 1–38.

———. 1904. “Über Die Matamorphose Einiger Phryganeiden Und Limnophiliden III.” *Acta Societas pro Fauna et Flora Fennica* 27 (2): 1–74.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics*, January, btu033. doi:10.1093/bioinformatics/btu033.

Stuart, Alan. 1955. “A Test for the Homogeneity of the Marginal Distributions in a Two-Way Classification.” doi:10.1093/biomet/42.3-4.412.

Sukatsheva, I. D. 1968. “Mesozoic Caddisflies (Trichoptera) from Transbaikalia.” *Paleontol. Zhurn* 2: 59–75.

Sukatsheva, I.D. 1980. “Evolution of the Caddisfly (Trichoptera) Larval Case Construction.” *Zhurnal Obschei Biologii* 41 (3): 457–69.

———. 1992. “New Fossil Representatives of Caddisflies from Mongolia.” *Joint Russian-Mongolian Paleontological Expedition, Transactions* 41: 111–17.

Tavaré, S. 1986. “Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.” *Lectures on Mathematics in the Life Sciences (American Mathematical Society)* 17: 57–86.

Van Dijk, Erwin L., Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. “Ten Years of next-Generation Sequencing Technology.” *Trends in Genetics* 30 (9): 418–26. doi:10.1016/j.tig.2014.07.001.

Whitfield, James B., and Karl M. Kjer. 2008. “Ancient Rapid Radiations of Insects: Challenges for Phylogenetic Analysis.” *Annual Review of Entomology* 53 (1): 449–72. doi:10.1146/annurev.ento.53.103106.093304.

Wiggins, Glenn B. 1956. “A Revision of the North American Caddisfly Genus Banksiola (Trichoptera: Phryganeidae).” *Contributions of the Royal Ontario Museum, Division of Zoology and Paleontology* 43: 1–12.

———. 1960a. “A Preliminary Systematic Study of the North American Larvae of the Caddisfly Family Phryganeidae (Trichoptera).” *Canadian Journal of Zoology* 38 (6): 1153–70.

———. 1960b. “The Unusual Pupal Mandibles in the Caddisfly Family Phryganeidae (Trichoptera).” *Canadian Entomologist* 96 (6): 449–57.

———. 1962. “A New Subfamily of Phryganeid Caddisflies from Western North America (Trichoptera: Phryganeidae).” *Canadian Journal of Zoology* 40 (5): 879–91.

———. 1972. “The Caddisfly Family Phryganeidae: Classification and Phylogeny for the World Fauna (Trichoptera). Symposium A. Systematics, Ecology, and Phylogeny of Odonata, Ephemeroptera, Plecoptera, and Trichoptera.” *Proceedings of the XIII International Congress of Entomology*, 342.

———. 1998. *The Caddisfly Family Phryganeidae*. 1st ed. University of Toronto Press, Scholarly Publishing Division.

Wiggins, Glenn B., and D.J. Larson. 1989. “Systematics and Biology for a New Nearctic Genus in the Caddisfly Family Phryganeidae (Trichoptera).” *Canadian Journal of Zoology* 67 (6): 1550–56.

Williams, David M. 2013. “Why Is Synedra Berolinensis so Hard to Classify? More on Monotypic Taxa.” *Phytotaxa* 127 (1): 113–27. doi:10.11646/phytotaxa.127.1.13.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. “SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads.” *Bioinformatics* 30 (12): 1660–66. doi:10.1093/bioinformatics/btu077.

Yang, Ziheng. 1994. “Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods.” *Journal of Molecular Evolution* 39 (3): 306–14. doi:10.1007/BF00160154.

———. 1996. “Among-Site Rate Variation and Its Impact on Phylogenetic Analyses.” *Trends in Ecology & Evolution* 11 (9): 367–72. doi:10.1016/0169-5347(96)10041-0.