Masthead Logo

Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2018-03-01

Molecular Evolution of Odonata Opsins, Odonata Phylogenomics and Detection of False Positive Sequence Homology Using Machine Learning

Anton Suvorov
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

BYU ScholarsArchive Citation

Suvorov, Anton, "Molecular Evolution of Odonata Opsins, Odonata Phylogenomics and Detection of False Positive Sequence Homology Using Machine Learning" (2018). *All Theses and Dissertations*. 7320.
https://scholarsarchive.byu.edu/etd/7320

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Molecular Evolution of Odonata Opsins, Odonata Phylogenomics and Detection of False Positive Sequence Homology Using Machine Learning

Anton Suvorov

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Seth M. Bybee, Chair
Mark J. Clement
Michael F. Whiting
Dan A. Ventura
Keith A. Crandall

Department of Biology

Brigham Young University

Copyright © 2018 Anton Suvorov

All Rights Reserved

ABSTRACT

Molecular Evolution of Odonata Opsins, Odonata Phylogenomics and Detection of False Positive Sequence Homology Using Machine Learning

Anton Suvorov
Department of Biology, BYU
Doctor of Philosophy

My dissertation comprises three related topics of evolutionary and computational biology, which correspond to the three Chapters. Chapter 1 focuses on tempo and mode of evolution in visual genes, namely opsins, via duplication events and subsequent molecular adaptation in Odonata (dragonflies and damselflies). Gene duplication plays a central role in adaptation to novel environments by providing new genetic material for functional divergence and evolution of biological complexity. Odonata have the largest opsin repertoire of any insect currently known. In particular our results suggest that both the blue sensitive (BS) and long-wave sensitive (LWS) opsin classes were subjected to strong positive selection that greatly weakens after multiple duplication events, a pattern that is consistent with the permanent heterozygote model. Due to the immense interspecific variation and duplicability potential of opsin genes among odonates, they represent a unique model system to test hypotheses regarding opsin gene duplication and diversification at the molecular level. Chapter 2 primarily focuses on reconstruction of the phylogenetic backbone of Odonata using RNA-seq data. In order to reconstruct the evolutionary history of Odonata, we performed comprehensive phylotranscriptomic analyses of 83 species covering 75% of all extant odonate families. Using maximum likelihood, Bayesian, coalescent-based and alignment free tree inference frameworks we were able to test, refine and resolve previously controversial relationships within the order. In particular, we confirmed the monophyly of Zygoptera, recovered Gomphidae and Petaluridae as sister groups with high confidence and identified Calopterygoidea as monophyletic. Fossil calibration coupled with diversification analyses provided insight into key events that influenced the evolution of Odonata. Specifically, we determined that there was a possible mass extinction of ancient odonate diversity during the P-Tr crisis and a single odonate lineage persisted following this extinction event. Lastly, Chapter 3 focuses on identification of erroneously assigned sequence homology using the intelligent agents of machine learning techniques. Accurate detection of homologous relationships of biological sequences (DNA or amino acid) amongst organisms is an important and often difficult task that is essential to various evolutionary studies, ranging from building phylogenies to predicting functional gene annotations. We developed biologically informative features that can be extracted from multiple sequence alignments of putative homologous genes (orthologs and paralogs) and further utilized in context of guided experimentation to verify false positive outcomes.

Keywords: molecular evolution, vision, insects, Bayesian modeling, phylogenetic inference, big data, next-generation sequencing, artificial intelligence, homology

ACKNOWLEDGEMENTS

First, I would like to thank members of my PhD committee, Profs. Seth Bybee, Mark Clement, Keith Crandall, Dan Ventura and Michael Whiting, who provided outstanding academic as well as personal support throughout my four years at BYU. I thank my advisor Seth who helped me become a better scientist, diversify my research interests and who gave me much needed inspiration and moral support.

None of my work would be possible without my previous academic advisors and their keen examples. I want to thank Galina Ryazanova and Ilya Zakharov (Moscow State University) who nurtured me by giving me essential skills and a foundation of knowledge. Francis Jiggins (Cambridge University) who turned me into a real-time PCR guru and helped me greatly strengthen my knowledge of evolutionary biology and genetics. Christian Schlötterer and Andreas Futschik (Institute of Population Genetics, Vienna) who got me interested in hypothesis-driven research, population genetics, bioinformatics and biostatistics. I want to acknowledge my colleagues, collaborators and peers: Daniel Fabian, Susanne Franssen, Nicola Palmieri, Ram Vinay Pandey, Carolin Kosiol, Maren Wellenreuther, Julia Hosp, Tatiana Galinskaya. These individuals have been inspiring and helped me with my research throughout my entire academic career.

Also, I want to thank my students Nick, Mitchell and Kyle, and my BYU colleagues Stanley, Milly, Nathan, Gavin, Sebastian, Paul, Haley, Robert, Gareth, Xia, Andrea, Andy, Joann, Sam, Rebecca and Yelena for their amazing ideas, creative thinking and constant help with my research.

I want to thank my Army leaders, LTC Forrest “Chip” Cook and 1LT Jaron Janson, my ROTC fellow Cadets and battle buddies who helped me develop leadership skills, discipline and resilience that in turn had an enormous positive influence on my academic career.

Last but not least, I want to thank my family and friends who were always there for me and for their genuine interest in my academic pursuits. Especially my two beautiful kids, Agnes and Alexander, my adoring wife, Kristina, and my supportive parents, Pavel and Olga, without whom my enthusiasm and passion for science would be non-existent.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF FIGURES

CHAPTER 1

Manuscript published in Molecular Ecology

Chapter 1

Opsins have evolved under the permanent heterozygote model: insights from phylotranscriptomics of Odonata

Anton Suvorov*§[1], Nicholas O. Jensen*[1], Camilla R. Sharkey[1], M. Stanley Fujimoto[2], Paul Bodily[2], Haley M. Cahill Wightman[1], T. Heath Ogden[3], Mark J. Clement[2] and Seth M. Bybee[1]

[1]Department of Biology, Brigham Young University, Provo, Utah 84602

[2]Computer Science Department, Brigham Young University, Provo, Utah 84602

[3]Department of Biology, Utah Valley University, Orem, Utah 84058

*Equal contributors

§Corresponding author

Abstract

Gene duplication plays a central role in adaptation to novel environments by providing new genetic material for functional divergence and evolution of biological complexity. Several evolutionary models have been proposed for gene duplication to explain how new gene copies are preserved by natural selection but these models have rarely been tested using empirical data. Opsin proteins, when combined with a chromophore, form a photopigment that is responsible for the absorption of light, the first step in the phototransduction cascade. Adaptive gene duplications have occurred many times within the animal opsins gene family, leading to novel wavelength sensitivities. Consequently, opsins are an attractive choice for the study of gene duplication evolutionary models. Odonata (dragonflies and damselflies) have the largest opsin

repertoire of any insect currently known. Additionally, there is tremendous variation in opsin copy number between species, particularly in the long wavelength sensitive (LWS) class. Using comprehensive phylotranscriptomic and statistical approaches we tested various evolutionary models of gene duplication. Our results suggest that both the blue sensitive (BS) and LWS opsin classes were subjected to strong positive selection that greatly weakens after multiple duplication events, a pattern that is consistent with the permanent heterozygote model. Due to the immense interspecific variation and duplicability potential of opsin genes among odonates, they represent a unique model system to test hypotheses regarding opsin gene duplication and diversification at the molecular level.

Introduction

Modern Odonata represent one of the more primitive groups of insects and one of the first lineages to evolve flight (Grimaldi & Engel 2005). They are also highly dependent on vision, particularly during the adult stage where other sensory modalities are lacking or poorly developed (e.g. olfaction (Rebora *et al.* 2012) and audition (Robert 2005)). The potential for color discrimination has been demonstrated physiologically in both damselflies (Huang *et al.* 2014) and dragonflies (Meinertzhagen *et al.* 1983b; Yang & Osorio 1991b) with adults possessing up to five spectrally distinct photoreceptors. A recent transcriptomics study (Futahashi *et al.* 2015) has shown that the genes responsible for mediating spectral sensitivity, namely opsins, have undergone significant expansions amongst odonate lineages. Odonates possess more opsin copies (11 – 30) than all other insects studied thus far and considerably more than the proposed tri-chromatic insect ancestor (Briscoe & Chittka 2001).

Opsins are transmembrane G-protein coupled receptors, which are covalently associated with a chromophore molecule forming a photopigment that is responsible for the absorption of light. The wavelength sensitivity of the photopigment is largely dictated by the structure of the opsin protein, which can be altered through changes in the amino acid sequence. The diversity of opsin sequences and thus photopigment wavelength sensitivities form the basis of colour vision in animals. The study of opsin proteins has garnered much attention from multiple fields, contributing to a comprehensive understanding of their structure, function and phylogeny. Due to the causal relationship between opsin structure and photopigment wavelength sensitivity, it is possible to explore the genetic basis of color vision and its evolution. Indeed, large phylogenetic studies have described much of the wide-scale variation in spectral sensitivity observed across the animal kingdom by tracing the losses and gains of photopigments (Feuda *et al.* 2012; Hering

et al. 2012; Rivera *et al.* 2010). Strikingly, although opsins are well characterized and there is a wealth of opsin sequence data, no studies have yet explored the evolutionary models that describe the process of opsin duplication, which is fundamental to the evolution of novel visual pigments and hence the underlying mechanisms that describe the variation in color vision systems we observe amongst animals today.

Gene duplications provide new genetic material for evolution, shaping its tempo and mode (Conrad & Antonarakis 2007; Hahn 2009; Han *et al.* 2009; Innan & Kondrashov 2010; Kondrashov 2012; Ohta 1989; Zhang 2003). New gene duplicates are usually lost due to the low probability of fixation in a population (Lynch *et al.* 2001). However, duplicates with a strong selective advantage have more potential to be preserved and increase in frequency within a population. Models of gene-duplication evolution have three key phases that may lead to copy maintenance: fixation, fate-determination and preservation (Innan and Kondrashov 2010). Each model can be tested for by examining the patterns of natural selection observed during the three different phases making it possible to hypothesize about the type of gene functionalization that has occurred.

Currently, 10 different models of gene-duplication evolution grouped into four categories (I-IV) (Table 1) are established in the literature (Innan and Kondrashov 2010). Category I models assume that a newly duplicated copy does not have a fitness advantage (or disadvantage), so that fixation is a neutral process. Depending on the model in this category, a new copy can evolve a novel function (neofunctionalization) or retain a part of the original ancestral function (subfunctionalization). A notable example is the evolution of duplicated *engrailed* genes of zebrafish, homeobox transcription factors responsible for the midbrain-hindbrain boundary formation, under the Duplication-Degeneration-Complementation (DDC) model (Force *et al.*

1999; Stoltzfus 1999). In this case, under relaxed functional constraints the complementary degenerate mutations most likely were fixed at random (i.e. with probability of $1/2N_e$), indicating evolution of duplicated genes *via* a neutral process. Category II models assume that duplication itself is adaptive and positive selection is exerted on a newly duplicated copy. A new copy usually retains its original function but can also be neofunctionalized. The evolution of new gene copies was accelerated by positive selection for both antimicrobial genes in *Drosophila* genomes (Sackton *et al.* 2007) and for amylase genes in humans (Perry *et al.* 2007). The models in category III are often characterized by multiple duplication events and lack the fate-determination phase since newly duplicated gene copies are immediately adaptive. Importantly, those models require high levels of allelic polymorphism created by strong positive selection during the pre-duplication phase whereas after duplication has occurred a new copy can be subjected to different selection modes. In category III, a functional divergence between gene duplicates results in neofunctionalization or subfunctionalization of a new copy. For example, acetylcholinesterase (*ace-1*), a gene responsible for insecticide resistance, in the mosquito *Culex pipiens* (Proulx & Phillips 2006; Spofford 1969) prevents further segregation of adaptive allelic combinations (i.e. heterozygote advantage (Sellis *et al.* 2011)) by fixing the resistance allele *ace-1*^{R} with susceptible copies through a duplication event. Such duplications conform to the permanent heterozygote model of Category III (Labbe *et al.* 2007). Category IV includes a single dosage balance model where fixation of new copies is determined by large-scale genetic aberrations such as large segmental or whole chromosome/genome duplications. Where dosage imbalance has deleterious effects, this model predicts that newly duplicated genes enter the preservation phase immediately and subsequently all copies are subjected to (relaxed) purifying selection. Both new and old copies retain their original function. An example of this model is

dosage-sensitive genes in yeast (*Saccharomyces cerevisiae*) (Papp *et al.* 2003) and mammals (Schuster-Bockler *et al.* 2010) that produce protein complexes with multiple subunits.

No explicit test of duplication models that explain opsin evolution among arthropods has been conducted to date. However, Yuan *et al.* (2010) showed that in *Heliconius* butterflies the rate of synonymous substitutions exceeded the rate of non-synonymous substitutions on the branch leading to one of two UV opsin clades (UVRh2). Although the authors did not suggest an evolutionary model for *Heliconius* opsin evolution, they were able to exclude neofunctionalization (Innan & Kondrashov 2010; Ohno 1970) a Category I model. The pivotal placement of odonates in the evolutionary history of insects and their reliance on vision makes them a premier system to explore the complex models of evolution and diversification of opsin genes that underpin their visual systems.

First, we reconstructed a robust odonate species tree that was necessary for downstream evolutionary analyses. Second, using codon-based maximum likelihood branch-site models of positive selection (Yang & Nielsen 2002; Zhang *et al.* 2005) together with our naive Bayesian "diffusion" model of positive selection, we found that all three opsin clades (UVS, BS and LWS) evolved under strong positive selection during the pre-duplication phase. More specifically, using evidence from 16 odonate species, the Bayesian model was able to show: (i) that opsin copies are being fixed after the first duplication event and (ii) future copies become almost immediately fixed as supported by a pattern of positive selection "weakening" very quickly throughout the opsin tree. Taken together, these results strongly suggest that odonate opsins evolve under the permanent heterozygote model.

Materials and Methods

Taxon sampling, library preparation and RNA-seq

All samples were taken from adult males, collected in the USA. The data set comprised 18 Odonata (dragonflies and damselflies, 16 species including two biological replicates for *Anax junius* and *Hetaerina americana*) and two outgroup Ephemeroptera (mayflies) specimens. Total RNA was extracted for each taxon from eye tissue using NucleoSpin columns (Clontech) and reverse-transcribed into cDNA libraries using the Illumina TruSeq RNA v2 sample preparation kit that both generates and amplifies full-length cDNAs. Prepped Ephemeroptera mRNA libraries were sequenced on an Illumina HiSeq 2000 producing 101 bp paired-end reads (2 × 101-bp) by the Microarray and Genomic Analysis Core Facility at the Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT, USA, while all Odonata preps were sequenced on a GaIIx producing 72 bp paired-end reads (2 × 72-bp) by the DNA sequencing center at Brigham Young University, Provo, UT, USA. The expected insert sizes were 150 bp and 280 bp respectively. Raw RNA-seq reads were deposited in the National Center for Biotechnology Information (NCBI), Sequence Read Archive with the accession numbers specified in Table S1 (Supporting information).

Read trimming and de novo transcriptome assembly

The read libraries were trimmed using the Mott algorithm implemented in PoPoolation (Kofler *et al.* 2011) with default parameters (minimum read length = 40, quality threshold = 20). For the assembly of the transcriptome contigs we used Trinity (Grabherr *et al.* 2011), currently the most accurate *de novo* assembler for RNA-seq data (Zhao *et al.* 2011), under the default parameters. Since we had biological replicates for *Hetaerina americana* and *Anax junius*, we

combined RNA-seq replicated libraries for these two species, trimmed them and assembled using the aforementioned tools with identical parameters into "representative" transcriptomes with higher coverage. Table S1 (Supporting information) contains general information about RNA-seq libraries and assemblies. Transcriptome assemblies are available at Dryad http://dx.doi.org/10.5061/dryad.pb5vv.

To identify putative coding sequences within the Trinity assemblies we used TransDecoder (http://transdecoder.github.io), the utility that identifies the longest open reading frames (ORFs) within each assembled DNA contig (for further details see Appendix S1, Supporting information). To reduce redundancy of the predicted protein collections caused by misassembly, we removed all identical DNA contigs from Trinity assemblies and their corresponding protein sequences from each proteome using CD-HIT (Fu *et al.* 2012).

Orthology assignment, cluster filtering and phylogenetic inference

To identify the best strategy to reconstruct a phylogenetic species tree we used different orthology detection algorithms implemented in InParanoid-MultiParanoid v4.1 (Alexeyenko *et al.* 2006; Remm *et al.* 2001), OrthoMCL (Li *et al.* 2003) and HaMStR v13.2.2 (Ebersberger *et al.* 2009) as well as phylogenetic approaches including, Maximum Likelihood: IQ-TREE (Nguyen *et al.* 2015), Bayesian: ExaBayes (Aberer *et al.* 2014), Alignment Free: Co-phylog (Yi & Jin 2013) and Coalescent-based: ASTRAL (Mirarab *et al.* 2014). We used the multiple sequence alignment (MSA) algorithm of MAFFT (Katoh *et al.* 2002) and alignment quality filtering procedures such as ALISCORE (Misof & Misof 2009) and machine learning false-positive homology detection (Fujimoto *et al.* 2016a) implemented in GOCleaner (Fujimoto *et al.* 2016b). All approaches of orthology assignment/phylogenetic reconstruction returned putative species trees that were concordant and generally topologically consistent with nearly identical

branch lengths, uniformly high bootstrap supports and posterior probabilities. Throughout the paper we used the ML tree estimated from OrthoMCL clusters as the most likely estimate of Odonata phylogeny. For further details see Appendix S1, S2 and Fig. S1 (Supporting information).

Dating

A Bayesian algorithm of MCMCTREE (Yang 2007) was implemented to estimate divergence times within Odonata with 11 fossil constraints Appendix S1 (Supporting information) using the ML topology inferred from 1:1 OrthoMCL isoforms removed clusters and WAG+Γ substation model . The analysis was run independently 1,000 times for 8×10^6 generations, logging every 50th generation and then discarding 25% as burn-in.

Detection, filtering and analyses of opsin sequences

Searches for putative opsins in the Odonata and Ephemeroptera proteomes were performed against database using insect visual opsin homologous group EOG8NKF98 from OrthoDB database v8 (Waterhouse *et al.* 2013). This included long-wavelength sensitive (LWS), blue-sensitive (BS) and ultraviolet-sensitive (UVS) opsin classes. Opsins within this group were aligned using MAFFT and converted into a profile Hidden Markov Model (pHMM) database using hmmbuild program of HMMER (Eddy 2011). Then we screened TransDecoder-predicted proteomes for opsins against pHMM database using hmmscan with an E value cutoff of 10^{-10}. The corresponding DNA opsin sequences were extracted from TransDecoder-predicted CDSs. Additionally, we used PIA (Speiser *et al.* 2014) with default parameters to identify opsins that could be missed by the HMMER search using raw Trinity transcriptomes. All redundant (identical) opsin sequences identified by both approaches were removed using CD-HIT. Non-visual opsins, confirmed by BLASTing against all known adult and larval odonate opsin

sequences (including RGR, pteropsin, arthropsin and Rh7 opsins) (Futahashi *et al.* 2015), were excluded since they have no duplicates. Further, to detect any artifact opsin contigs assembled by Trinity, we mapped trimmed reads against Trinity transcriptomes and calculated expression values using RSEM (Li & Dewey 2011). We then log-transformed FPKM values and fitted a skewed normal distribution, which was used to identify lowly expressed opsins using 0.05 as our rejection level (opsin genes with FPKM < 30 for Odonata and < 1 for Ephemeroptera were excluded). Also, we visually inspected read – opsin contig alignments to check for potential chimeric sequences, assuming that chimeric contigs would produce non-uniform read coverage. For a complete list of opsins see Table S2 (Supporting information).

Putative Odonata and Ephemeroptera opsin protein sequences including homologous opsin sequences from mollusks (outgroup) were aligned using COBALT (Papadopoulos & Agarwala 2007), manually checked, 3'-UTR regions removed and were then back-translated into DNA sequences (Alignment 1). Using COBALT with a conserved domain database we were able to derive an accurate structural opsin alignment that was necessary for further site-specific evolutionary analyses. The opsin gene tree was estimated using IQ-TREE with the best-fit substitution model and with 10,000 UFBoot iterations to assess nodal support. Also, we created two additional datasets and performed all the analyses described below to ensure robustness of our selection inference. To create the first dataset, we excluded all partial opsin sequences (< 1056 bp and < 7 transmembrane domains) from our dataset (Alignment 2). To create the second dataset, we combined our opsin sequences with those identified from 12 species in (Futahashi *et al.* 2015) (Alignment 3). All three alignments are available at Dryad http://dx.doi.org/10.5061/dryad.pb5vv.

Tests for possible episodic positive selection operating on opsins were performed in PAML v4 (Yang 2007). Using branch-site new model A we tested whether the ancestral branches of LWS, BS and UVS opsins as well as their sites were affected by positive selection. In order to map positively selected sites in opsin protein domains we performed ancestral sequence reconstruction using the empirical Bayes approach (Yang *et al.* 1995) implemented in PAML and utilized inferred protein sequences to model structural domains using I-TASSER v4.3 (Yang & Zhang 2015) under default parameters using squid rhodopsin as a template (PDB model 2Z73A) (Murakami & Kouyama 2008). Additionally, using site models M0, M1a, M2a, M3, M7 and M8 we separately tested LWS, BS and UVS opsin classes for positive selection. For both site and site-branch models, the log-likelihood of each competing model was compared against the null model of fixed $\omega = 1$ (no selection) with the Likelihood Ratio Test (LRT) using χ^2 distributions with the appropriate degrees of freedom. To avoid a model getting trapped in a local optimum we ran analyses at least tree times specifying initial ω values at 0.1, 1 and 2. The Bayes empirical Bayes (BEB) (Yang *et al.* 2005) procedure was then used to calculate posterior probabilities for the site classes. In addition, using our Bayesian "diffusion" model of selection, we estimated probabilities of sites in the opsin copies to be under a certain selection mode before and after duplication events for LWS and BS classes.

Ancestral reconstruction of opsin turnover events was carried out using CAFE v3 (De Bie *et al.* 2006; Han *et al.* 2013). The opsin class contractions and expansions were estimated with the stochastic model of gene gains (birth) and losses (death) along the given ultrametric species tree (with collapsed *Hetaerina* and *Anax* branches). In order to ensure convergence, the analysis was repeated 1,000 times assuming a single global birth parameter and assuming birth variation between Anisoptera, Zygoptera and Epemeroptera lineages. A likelihood ratio test statistic was

then calculated as 2 × (best log-likelihood of global birth model – best log-likelihood of three-birth model) and compared against a null distribution of 2×10^4 simulated likelihood ratios using the -genefamily CAFE command (approximate LRT). Comparing the empirical and estimated number of copies with the null model of random birth-death process assessed global significance of each opsin class changes ($P \leq 0.05$) as well as the significance of individual branch gains/losses using the Viterbi algorithm.

Simulations

Old and long ancestral opsin branches may have been highly saturated, leading to the detection of false-positive positive selection (a type I error) by branch-site models. To make sure this was not the case, we performed simulations under the neutral branch-site model and tested ancestral lineages corresponding to UVS, BS and LWS for presence of positive selection. We simulated 1,000 datasets under the null model of the branch-site test (Yang & Nielsen 2002) using the evolverNSbranchsites program in PAML. Codon frequencies were estimated from the opsin MSA using PAML, then the 400 codons were drawn randomly from four site classes (see Table 1 in (Zhang, et al. 2005)) in proportions $p_0 = 0.6$, $p_1 = 0.2$, $p_{2a} = 0.15$, $p_{2b} = 0.05$. The ω ratios were specified for each class as follows: ($\omega_0 = 0.2$, $\omega_0 = 0.2$), ($\omega_1 = 1$, $\omega_1 = 1$), ($\omega_0 = 0.2$, $\omega_2 = 1$) and ($\omega_1 = 1$, $\omega_2 = 1$), where ω value in each pair corresponds to background and foreground branches of the opsin tree (Fig. 1B and Fig. S2, Supporting information). The branches UVS, BS and LWS were considered as foreground branches. Each simulated replicate was analyzed under branch-site model A (null: $\omega_1 = 1$ fixed vs. A: $\omega_1 \geq 1$ estimated). Test statistic 2ΔLnL was recorded for each comparison. As expected from the neutral model (Zhang *et al.* 2005), each branch test statistics 2ΔLnL = 2 × (LnL A – LnL Null) followed a 50:50 mixture distribution at point mass 0 and the χ_1^2 with the approximately 50% proportion of 2ΔLnL

= 0 (proportion of zeros for UVS:0.413, BS:0.410, LWS: 0.497) with a very low false-positive rate (Yang & dos Reis 2011) at a 0.05 level of significance (false-positive rate for UVS:0.049, BS:0.048, LWS: 0.0191) using strict branch-site test.

Naive Bayesian "diffusion" model of selection

We developed a simple 1-Level hierarchical Bayesian model to estimate the probability of a site being under purifying selection ($\omega < 1$), neutrality ($\omega = 1$) or positive selection ($\omega > 1$) before and after opsin duplication events. The number of PAML-predicted sites in each selection class under branch-site model (Zhang *et al.* 2005) across all species was used as evidence in the model. We explicitly binned LWS branches of an opsin tree inferred from each species (using all UVS sequences as an outgroup) into three categories, namely 0-before duplication, 1-duplicated branches, 2-terminal branches. For BS we used the same partitioning scheme.

In order to define our model we introduce a following notation:

$\boldsymbol{X} = (X_{1(\omega<1)}, X_{2(\omega=1)}, X_{3(\omega>1)})$ = inferred number of sites under purifying selection, neutrality and positive selection respectively,

$\boldsymbol{p} = (p_1, p_2, p_3)$ = parameter vector that specifies a probability of a site to belong to one of the aforementioned selection classes (K=3),

α_k = a parameter that specifies number of sites (successes) in each class K,

$\hat{a}_k$ and b = parameters that completely define hyperprior distributions for each α_k,

L = average length of an opsin protein sequence,

i = 1, … N, number of branches in each duplication category (bin). $\hat{a}_k$ were calculated from the data as the average selection class size length multiplied by b. By doing so we centered each hyperprior distribution on the average class size using the notion that the first moment of a

Gamma distribution is equal to $\frac{a}{b}$. Apparently $L = \alpha_1 + \alpha_2 + \alpha_3$, thus we constrained α_3 to $L - \alpha_1 - \alpha_2$.

Model specification:

$$\boldsymbol{X} \mid \boldsymbol{p}, n \sim \text{Multinomial}(\boldsymbol{p}, n)$$

$$\boldsymbol{p} \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\alpha_1 \sim \text{Gamma}(\hat{a}_1\ ,\ \text{b} = 0.5)$$

$$\alpha_2 \sim \text{Gamma}(\hat{a}_2\ ,\ \text{b})$$

$$\alpha_3 \sim \text{Gamma}(\hat{a}_3\ ,\ \text{b})$$

For Bayesian inference we obtain the posterior distribution up to a multiplicative constant as following:

$$P(\boldsymbol{p}, \boldsymbol{\alpha} \mid \boldsymbol{X}, L, \hat{a}_1, \hat{a}_2, \hat{a}_3, \text{b})$$

$$\propto \left\{\prod_{k=1}^{3}\prod_{i=1}^{N} \text{Multinomial}(X_{ki} \mid p_k,)\right\}\{\text{Dirichlet}(\boldsymbol{\alpha} \mid L)\}\left\{\prod_{k=1}^{3} \text{Gamma}(\hat{a}_k\ ,\ \text{b})\right\}$$

Complete (full) conditionals for Metropolis-within-Gibbs sampling:

$$P(\boldsymbol{p}|all) \propto p_1^{\sum X_{1(\omega<1)}} p_2^{\sum X_{2(\omega=1)}} p_3^{\sum X_{3(\omega>1)}} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{L-\alpha_1-\alpha_2-1}$$

$$\propto \text{Dirichlet}\Big(\sum X_{1i(\omega<1)} + \alpha_1 - 1, \sum X_{2i(\omega=1)} + \alpha_2 - 1, \sum X_{3i(\omega>1)} + L - \alpha_1 - \alpha_2 - 1\Big)$$

$$\log\big(P(\alpha_1|all)\big) \propto \log\left(\frac{1}{\Gamma(\alpha_1)\Gamma(L-\alpha_1-\alpha_2)} p_1^{\alpha_1} p_3^{-\alpha_1} \alpha_1^{\hat{a}_1-1} e^{-\text{b}\alpha_1} (L-\alpha_1-\alpha_2)^{\hat{a}_3-1} e^{\text{b}\alpha_1}\right)$$

$$\propto -\big[\log\big(\Gamma(\alpha_1)\big) + \log\big(\Gamma(L-\alpha_1-\alpha_2)\big)\big] + \alpha_1 \log\frac{p_1}{p_3} + (\hat{a}_1 - 1)\log\alpha_1$$

$$+ (\hat{a}_3 - 1)\log(L-\alpha_1-\alpha_2)$$

$$\log\big(P(\alpha_2|all)\big) \propto \log\Big(\frac{1}{\Gamma(\alpha_2)\Gamma(L-\alpha_1-\alpha_2)} p_2^{\alpha_2} p_3^{-\alpha_2} \alpha_2^{\hat{a}_2-1} e^{-b\alpha_2} (L-\alpha_1-\alpha_2)^{\hat{a}_3-1} e^{b\alpha_2}\Big)$$

$$\propto -\big[\log\big(\Gamma(\alpha_2)\big) + \log\big(\Gamma(L-\alpha_1-\alpha_2)\big)\big] + \alpha_2 \log\frac{p_2}{p_3} + (\hat{a}_2 - 1)\log\alpha_2$$

$$+ (\hat{a}_3 - 1)\log(L-\alpha_1-\alpha_2)$$

Since complete conditionals of the parameters α_1 and α_2 have no closed form, we sampled them using the Metropolis algorithm with the normal proposal $u \sim \mathrm{N}(0, 0.5)$. Convergence was verified for all parameters in all models.

All statistical analyses were implemented in the R programming language. The R Bayesian model script is available at Dryad http://dx.doi.org/10.5061/dryad.pb5vv.

Results

Phylogenetic Inference

Using different combinations of orthology and phylogenetic algorithms, all putative species trees were topologically congruent with high bootstrap supports/posterior probabilities and consistent branch lengths (Figs. S3-S5, Supporting information). We used the ML tree estimated from the protein supermatrix of 1:1 orthologous 770 OrthoMCL gene clusters (285,648 sites) as the most likely estimate of Odonata phylogeny (see Appendix S1, S2 and Fig. S3 (Supporting information) for more details). These gene clusters represent a group of genes that are consistently expressed in odonate heads and presented as a single copy in each species.

Dynamics of opsin gene gains/losses

We investigated the evolutionary forces that operate on expressed odonate opsin genes. First we identified tremendous expansion of visual-opsin genes within the adult stage of Odonata, which confirm previous findings of Futahashi *et al.* (2015). They characterized the large expansion of the opsin gene family mainly in the LWS class, and in the suborder of Anisoptera (14 to 30 copies), suggesting a potentially complex color vision system (Futahashi *et al.* 2015). Our total estimates of opsin copy number range from 10 to 30 in Anisoptera and from 9 to 22 in Zygoptera (Fig. 1A). The common ancestor of Odonata was reconstructed to have: 1 ultra-violet sensitive (UVS), 2 blue sensitive (BS), 10 long-wavelength sensitive (LWS). A birth-death stochastic model implemented in CAFE (De Bie *et al.* 2006; Han *et al.* 2013) revealed variation in patterns of ancestral opsin class sizes between and within both suborders (Fig. 1A). The rate of opsin gains (birth parameter λ) differed significantly between Anisoptera and Zygoptera (approximate LRT, $P < 10^{-10}$) with more pronounced expansions observed in Anisoptera, i.e. $\lambda_{Anisoptera} (= 0.0038) > \lambda_{Zygoptera} (= 0.0029)$.

When LWS and BS opsin classes were tested separately for potentially significant turnovers we found that only the LWS class underwent globally significant expansions (Family-wide $P = 0.0006$) along the phylogeny. Within the BS class there was less variation in turnover (Family-wide $P = 0.963$). Six species experienced significant turnover of the LWS gene copies at the tree terminal branches (Viterbi method with the randomly generated likelihood distribution, $P \leq 0.05$; Fig. 1A, black stars), i.e. the number of gains/losses was considerable compared to the ancestral state.

Analyses of opsin sequence evolution

To test for potential positive selection operating on ancestral opsin lineages (Fig. 1B and Fig. S2, Supporting information) we took a maximum likelihood approach of branch-site models and performed ancestral reconstruction of UVS, BS and LWS (Yang 2007). We also predicted putative 3D protein models and chromophore binding site using I-TASSER (Yang & Zhang 2015). Odonata ancestral opsin lineages of UVS, BS and LWS classes as well as different sites (Fig. 2), were subjected to strong positive (diversifying) selection i.e. $\omega = dN/dS > 1$ (dN, number of non-synonymous substitutions per non-synonymous site; dS, number of synonymous substitutions per synonymous site) (LRT, $P = 0.00368$, Table 2). These observations were also confirmed using only full length opsin sequences as well as a combined dataset with all previously published odonate opsin sequences from 12 additional species (Futahashi *et al.* 2015) (Tables S3-S4 and Figs. S6-S8, Supporting information).

To investigate ongoing selection pressures acting on the three opsin classes, we used site codon models of molecular evolution implemented in PAML. Comparison of M3 vs. M0 site models suggest that selection mode (ω), i.e. heterogeneity in ω (Wong *et al.* 2004), varies among the sites of each opsin class, which is expected for functional protein coding genes (LRT, $P < 10^{-10}$, Table 3). However, the selection tests M2a vs. M1a and M8 vs. M7 showed no evidence for substantial positive selection for UVS, BS opsin classes, whereas M2a vs. M1a and M8 vs. M7 comparisons were found to be significant for LWS (LRT, $P < 10^{-5}$, Table 3 and Tables S5-S6, Supporting information).

To summarize selection patterns throughout opsin evolutionary history among species we used our Bayesian "diffusion" model of selection to identify the probability of a site under positive selection during the pre- and postduplicational phase for BS as well as LWS odonate opsins. We

found strong positive selection during only the preduplicational phase. The intensity of positive selection gradually decreases throughout the tree moving toward the terminal branches.

Discussion

Variation in opsin turnover rates between opsin classes as well as odonate suborders

According to our dated phylogeny, the majority of LWS turnover happened throughout the Cenozoic (65.5 MYA to present) and Cretaceous (145.4 to 65.5 MYA) of the Mesozoic (Fig. 1A). The expansion of angiosperm (flowering) plants during this geological time period initiated diversification of herbivorous insects (Grimaldi & Engel 2005), which were a food source for exclusively carnivorous Odonata. Thus, new predatory tactics may have been developed by Odonata and would have required visual enhancements, such as broadening visual sensitivity and increasing spectral discrimination. A difference in the number of opsin copies expressed during the adult stage between both suborders may be a result of adaptive evolution of visual systems related to flight and other behavioral or ecological strategies that differ between them (Sherk 1978). For example, many of the most visually complex Anisoptera (e.g., Aeshnidae and Libellulidae) are almost constantly in flight seeking prey and mates, while Zygoptera are generally perched and have targeted bouts of flight to intercept potential mates, competitors or prey. Such differences may serve to explain the opsin copy number variation between suborders, however more research is needed to address these ideas directly.

Comparing opsin spectral classes, signatures of positive selection have been only found within LWS opsin class using PAML site models (Table 3 and Tables S5-S6, Supporting information). Additionally, pairwise comparisons of mean ωs between odonate UVS, BS and LWS classes and all insect opsins also suggest that in general, Odonata LWS opsins are currently

evolving under relaxed selective pressure (Wilcoxon rank sum test, $P < 10^{-10}$, Fig. 3A and Fig. S9, Supporting information). These findings may explain the extraordinary LWS opsin expansions (gains) where relaxation of purifying selection and/or positive selection accelerate evolution and increase the amount of divergence (Arbiza *et al.* 2006; Ho-Huu *et al.* 2012).

Ancestral opsin chromophore binding pocket sites were not under positive selection

It is widely accepted that the spectral sensitivity of a visual pigment is determined by the composition of the light-receptive chromophore and the amino acid sequence of the opsin protein, specifically within the chromophore binding pocket (Asenjo *et al.* 1994; Merbs & Nathans 1993; Sun *et al.* 1997). The majority of positively selected sites in ancestral UVS, BS and LWS opsins were distributed away from the binding pocket, within the α-helices or loop regions (Fig. 2 and Fig. S8, Supporting information). Although mutations at these sites may affect the spectral tuning of the photopigment (Bowmaker 2008; Yokoyama 2008), it is more likely that they affect other characteristics of the opsin protein, such as stability or membrane binding (Dasmeh *et al.* 2013, 2014).

Selection patterns suggest the permanent heterozygote model of opsin gene evolution via duplication

Visual opsins belong to a multigene family that experienced various gene turnovers (mostly through gains) throughout the metazoan evolutionary history (Feuda *et al.* 2012; Henze & Oakley 2015; Porter *et al.* 2012; Rivera *et al.* 2010). However little is known about the evolutionary mechanisms that lead to opsin divergence and maintenance of duplicated copies, especially where extraordinary gains have occurred, like in Odonata.

Duplication of genetic material has a tremendous impact on organismal adaptation, rates of biological complexity evolution and diversification and can even reduce the probability of extinction (Crow *et al.* 2006; Donoghue & Purnell 2005; Kondrashov 2012; Lipinski *et al.* 2011; Qian & Zhang 2014). The biological advantage of extra genetic material can be seen in such processes as a beneficial increase in gene dosage (Qian & Zhang 2008), protection against deleterious mutations (Hsiao & Vitkup 2008), evolution of new functions (neofunctionalization) under changing environmental pressures (Lynch 2007) and others. In order to avoid pseudogenization, newly duplicated genes are expected to have positive contribution to fitness (Clark 1994) and their fixation is then determined by different selection forces (Innan & Kondrashov 2010; Kondrashov *et al.* 2002; Wagner 2002; Zhang *et al.* 1998).

The first study to explore insect opsin evolutionary dynamics (Yuan *et al.* 2010) implemented a small-sample method (a Fisher exact test-based statistical approach that compares proportion of synonymous and non-synonymous substitutions detected on a pre-duplication and all descendent branches (Zhang *et al.* 1997)) to test for episodic evolution on the branch leading to UVRh2 opsin in *Heliconius* butterflies. It has been shown that positive selection along the branch leading to UVRh2 followed purifying selection on proceeding branches. This method, however, lacks the possibility to summarize evolutionary patterns that occur in parallel in different species. Additionally, it is unable to incorporate uncertainty about the inferred number of synonymous and non-synonymous substitutions into a probabilistic model. Thus, we developed a novel flexible Bayesian "diffusion" model of selection, which predicts the probability of site-specific selection before and after gene duplication across an opsin gene tree (see Materials and Methods). By taking evidence (i.e. number of positively selected sites before and after opsin duplication events) our model aims to generalize evolutionary patterns that were

observed independently in odonate species. In particular, it shows that after the first duplication event, adaptive evolution quickly shifts toward purifying selection against non-synonymous substitutions (indicated by the decreased number of positively selected sites). Theoretically, this process gradually leads to eventual fixation of novel opsin copies in a genome (Fig. 3B and Fig. S10, Supporting information). Such strong episodic positive selection acting on alleles during the pre-duplication phase likely promoted maintenance of divergent opsin alleles within ancestral odonate populations (especially within the LWS class). From this genetic variation, when adaptive allelic combinations (heterozygote advantage) of opsins came together through duplication events, immediate fixation of the combinations followed (Spofford 1969). The permanent fixation of opsin duplicates in the population may have arisen *via* recombination (unequal crossing over) (Ohno 1970), thus minimizing the effect of segregation load (i.e., losing beneficial genetic associations generated by selection (Haag & Roze 2007)) and hence maximizing mean fitness (Hahn 2009). Additionally, these multiple opsin duplications may produce canalizing effects to counterbalance fitness load caused by segregation, mutations and/or environmental perturbations, thus increasing phenotypic robustness (Proulx & Phillips 2005).

During the pre-duplication period, positive selection is maximized but greatly weakens after duplication events where purifying selection is predominant (Fig. S10, Supporting information). We hypothesize that Odonata opsins evolved under DDC (Category I), the specialization (Category I) or the permanent heterozygote (Category III) models, since each supports the subfunctionalization of new copies, the probable functional opsin evolutionary trajectory (Lord *et al.* 2016; Spady *et al.* 2006). Nevertheless, since we detected strong positive selection in the LWS and BS opsins during the pre-duplication phase only the three models in Category III (Table 1) are possible to explain such patterns of molecular evolution: permanent

heterozygote (Proulx & Phillips 2006) (aka. segregation avoidance (Hahn 2009)), adaptive radiation (Francino 2005) or multiallelic diversifying selection (Penn *et al.* 2002). All three models lack a fate-determination phase and assume quick fixation of new gene duplicates (Innan and Kondrashov 2010). Each model is determined by selective patterns exerted on gene duplicates (Table 4). We note that after the first duplication events the probability of a site being under positive selection drops down significantly ($P \sim 0.22 \rightarrow P \sim 0.051$ for BS and $P \sim 0.378 \rightarrow P \sim 0.028$ for LWS, fig. 3B), however this probability is still larger when compared to that of terminal branches ($P \sim 0.051$ vs. $P \sim 0.024$ for BS and $P \sim 0.028$ vs. $P \sim 0.015$ for LWS, fig. 3B). These shifts indicate that even being under higher selection pressure following the first duplication, opsins have more freedom (relaxation of purifying selection) to evolve. Another explanation of the observed selection trajectories would be that the branch-site model is unable to detect short episodic positive selection that likely operated during multiple and rapid opsin duplication bursts.

Overall, these observed signatures of selection (positive followed by purifying) especially after the first duplication events for both LWS and BS classes suggest that odonate opsins evolve under the permanent heterozygote model (Proulx & Phillips 2006). Interestingly, the *ace-1* locus in the mosquito *Culex pipiens* has also been proposed to be evolving under the permanent heterozygote model. Two paralogs (which were former alleles) differed only at one amino acid site (Labbe *et al.* 2007) but this difference was sufficient to make the mosquito less susceptible to insecticide pressure. Also, the evolution of gene duplication under a permanent heterozygote model is the most likely mechanism proposed for the major histocompatibility complex (MHC) in mammals (Demuth *et al.* 2006). Both examples are characterized by high rates of gene duplication in a short time period (remarkably, < 40 years was required to acquire new duplicates

in mosquito), a pattern also observed for opsins in Odonata. Functional partitioning for opsin genes is a change in the amino acid sequence that leads to a subsequent change in the function of the visual pigment, such as a shift in its peak sensitivity. It is not known if odonate opsin copies found in this study exhibit functional partitioning, however, the permanent heterozygote model itself predicts subfunctionalization of opsin copies (Hahn 2009; Innan & Kondrashov 2010). Physiological evidence suggests there are more LWS opsin copies than required to explain the spectral sensitivities of odonate LWS photoreceptors (Meinertzhagen *et al.* 1983a; Yang & Osorio 1991a). However, there may be other functional differences between opsin copies such as regeneration, stability and membrane binding that would induce phenotypic invariance and fitness advantage.

In conclusion, the large expansion of opsin genes among odonates provides an exceptional system for testing hypotheses regarding the evolution of opsin genes and large-scale gene duplication, generally. Using transcriptomic data we were able to identify expressed opsin copies from across the order and demonstrate that odonate BS and LWS opsins evolve under the permanent heterozygote model, which suggests that opsins have undergone subfunctionalization. Future research with denser taxon sampling and genomic data are needed to further understand the mechanistic models of opsin duplication and their functional importance to opsin evolution in complex systems like Odonata and other arthropod groups.

Acknowledgements

We thank Gavin Martin and Nathan Lord for the generation of sequence data, Phill Watts and Keith Crandall for their critical comments and Rebecca Plimpton for her assistance with 3D protein modeling. We are grateful to Ziheng Yang for his valuable recommendations for

simulation analysis and comments on positive selection results. We also thank the Fulton Supercomputer lab (BYU) for assistance. This work was supported by the NSF grant (SMB; DEB-1265714) and by a BYU Graduate Research Fellowship (HMCW). The Associate Editor and referees provided insightful comments on earlier versions of the manuscript. None of the co-authors have declared a conflict of interest.

AS, NOJ and SMB designed the study. AS developed the Bayesian model, performed statistical and PAML analyses. CRS performed opsin 3D modeling. Phylogenetic inference was performed by NOJ with help from MSF, PB, THO and MJC. HMCW provided the dated phylogenetic tree. AS, SMB and CRS wrote the manuscript.

Data accessibility

Raw RNA-seq read libraries have been deposited in GenBank (for SRA IDs see Table S1, Supporting information). Opsin alignments (Alignment 1, Alignment 2 and Alignment 3), Bayesian model script and assembled trenscriptomes have been deposited in Dryad http://dx.doi.org/10.5061/dryad.pb5vv.

References

Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol Biol Evol* **31**, 2553-2556.

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15.

Arbiza L, Dopazo J, Dopazo H (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* **2**, e38.

Asenjo AB, Rim J, Oprian DD (1994) Molecular determinants of human red/green color discrimination. *Neuron* **12**, 1131-1138.

Bowmaker JK (2008) Evolution of vertebrate visual pigments. *Vision Res* **48**, 2022-2041.

Briscoe AD, Chittka L (2001) The evolution of color vision in insects. *Annu Rev Entomol* **46**, 471-510.

Clark AG (1994) Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci U S A* **91**, 2950-2954.

Conrad B, Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* **8**, 17-35.

Crow KD, Wagner GP, Investigators ST-NY (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* **23**, 887-892.

Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI (2013) Positively selected sites in cetacean myoglobins contribute to protein stability. *PLoS Comput Biol* **9**, e1002929.

Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI (2014) The influence of selection for protein stability on dN/dS estimations. *Genome Biol Evol* **6**, 2956-2967.

De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. *PLoS One* **1**, e85.

Donoghue PC, Purnell MA (2005) Genome duplication, extinction and vertebrate evolution. *Trends Ecol Evol* **20**, 312-319.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**, 157.

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195.

Feuda R, Hamilton SC, McInerney JO, Pisani D (2012) Metazoan opsin evolution reveals a simple route to animal vision. *Proc Natl Acad Sci U S A* **109**, 18868-18872.

Force A, Lynch M, Pickett FB*, et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-1545.

Francino MP (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**, 573-577.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152.

Fujimoto MS, Suvorov A, Jensen NO, Clement MJ, Bybee SM (2016a) Detecting false positive sequence homology: a machine learning approach. *BMC Bioinformatics* **17**, 101.

Fujimoto MS, Suvorov A, Jensen NO*, et al.* (2016b) The OGCleaner: filtering false-positive homology clusters. *Bioinformatics*.

Futahashi R, Kawahara-Miki R, Kinoshita M*, et al.* (2015) Extraordinary diversity of visual opsin genes in dragonflies. *Proc Natl Acad Sci U S A* **112**, E1247-1256.

Grabherr MG, Haas BJ, Yassour M*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652.

Grimaldi DA, Engel MS (2005) *Evolution of the insects* Cambridge University Press, Cambridge, UK ; New York, NY.

Haag CR, Roze D (2007) Genetic load in sexual and asexual diploids: segregation, dominance and genetic drift. *Genetics* **176**, 1663-1678.

Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* **100**, 605-617.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Res* **19**, 859-867.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30**, 1987-1997.

Henze MJ, Oakley TH (2015) The Dynamic Evolutionary History of Pancrustacean Eyes and Opsins. *Integr Comp Biol*.

Hering L, Henze MJ, Kohler M*, et al.* (2012) Opsins in onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. *Mol Biol Evol* **29**, 3451-3458.

Ho-Huu J, Ronfort J, De Mita S*, et al.* (2012) Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.). *BMC Evol Biol* **12**, 195.

Hsiao TL, Vitkup D (2008) Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet* **4**, e1000014.

Huang SC, Chiou TH, Marshall J, Reinhard J (2014) Spectral Sensitivities and Color Signals in a Polymorphic Damselfly. *PLoS One* **9**.

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**, 97-108.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066.

Kofler R, Orozco-terWengel P, De Maio N*, et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**, e15925.

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci* **279**, 5048-5057.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biol* **3**, RESEARCH0008.

Labbe P, Berthomieu A, Berticat C*, et al.* (2007) Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito Culex pipiens. *Mol Biol Evol* **24**, 1056-1067.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Lipinski KJ, Farslow JC, Fitzpatrick KA*, et al.* (2011) High spontaneous rate of gene duplication in Caenorhabditis elegans. *Curr Biol* **21**, 306-310.

Lord NP, Plimpton RL, Sharkey CR*, et al.* (2016) A cure for the blues: opsin duplication and subfunctionalization for short-wavelength sensitivity in jewel beetles (Coleoptera: Buprestidae). *BMC Evol Biol* **16**, 107.

Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**, 1789-1804.

Lynch VJ (2007) Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol* **7**, 2.

Meinertzhagen IA, Menzel R, Kahle G (1983a) The identification of spectral receptor types in the retina and lamina of the dragonfly Sympetrum rubicundulum. *Journal of Comparative Physiology* **151**, 295-310.

Meinertzhagen IA, Menzel R, Kahle G (1983b) The Identification of Spectral Receptor Types in the Retina and Lamina of the Dragonfly Sympetrum-Ribicundulum. *Journal of Comparative Physiology* **151**, 295-310.

Merbs SL, Nathans J (1993) Role of hydroxyl-bearing amino acids in differentially tuning the absorption spectra of the human red and green cone pigments. *Photochem Photobiol* **58**, 706-710.

Mirarab S, Reaz R, Bayzid MS*, et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-548.

Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* **58**, 21-34.

Murakami M, Kouyama T (2008) Crystal structure of squid rhodopsin. *Nature* **453**, 363-367.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274.

Ohno S (1970) *Evolution by gene duplication* Allen & Unwin;Springer-Verlag, London New York,.

Ohta T (1989) Role of gene duplication in evolution. *Genome* **31**, 304-310.

Papadopoulos JS, Agarwala R (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **23**, 1073-1079.

Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194-197.

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A* **99**, 11260-11264.

Perry GH, Dominy NJ, Claw KG*, et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-1260.

Porter ML, Blasic JR, Bok MJ*, et al.* (2012) Shedding new light on opsin evolution. *Proc Biol Sci* **279**, 3-14.

Proulx SR, Phillips PC (2005) The opportunity for canalization and the evolution of genetic networks. *American Naturalist* **165**, 147-162.

Proulx SR, Phillips PC (2006) Allelic divergence precedes and promotes gene duplication. *Evolution* **60**, 881-892.

Qian W, Zhang J (2008) Gene dosage and gene duplicability. *Genetics* **179**, 2319-2324.

Qian W, Zhang J (2014) Genomic evidence for adaptation by gene duplication. *Genome Res* **24**, 1356-1362.

Rebora M, Salerno G, Piersanti S, Dell'Otto A, Gaino E (2012) Olfaction in dragonflies: Electrophysiological evidence. *J Insect Physiol* **58**, 270-277.

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052.

Rivera AS, Pankey MS, Plachetzki DC*, et al.* (2010) Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC Evol Biol* **10**, 123.

Robert D (2005) Directional hearing in insects. *Springer Handbook of Auditory Research* **25**, 6-35.

Sackton TB, Lazzaro BP, Schlenke TA*, et al.* (2007) Dynamic evolution of the innate immune system in Drosophila. *Nat Genet* **39**, 1461-1468.

Schuster-Bockler B, Conrad D, Bateman A (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One* **5**, e9474.

Sellis D, Callahan BJ, Petrov DA, Messer PW (2011) Heterozygote advantage as a natural consequence of adaptation in diploids. *Proc Natl Acad Sci U S A* **108**, 20666-20671.

Sherk TE (1978) Development of the compound eyes of dragonflies (Odonata). III. Adult compound eyes. *J Exp Zool* **203**, 61-80.

Spady TC, Parry JW, Robinson PR*, et al.* (2006) Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Mol Biol Evol* **23**, 1538-1547.

Speiser DI, Pankey MS, Zaharoff AK*, et al.* (2014) Using phylogenetically-informed annotation (PIA) to search for light-interacting genes in transcriptomes from non-model organisms. *BMC Bioinformatics* **15**, 350.

Spofford JB (1969) Heterosis and Evolution of Duplications. *American Naturalist* **103**, 407-&.

Stoltzfus A (1999) On the possibility of constructive neutral evolution. *J Mol Evol* **49**, 169-181.

Sun H, Macke JP, Nathans J (1997) Mechanisms of spectral tuning in the mouse green cone pigment. *Proc Natl Acad Sci U S A* **94**, 8860-8865.

Wagner A (2002) Selection and gene duplication: a view from the genome. *Genome Biol* **3**, reviews1012.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41**, D358-365.

Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**, 1041-1051.

Yang EC, Osorio D (1991a) Spectral sensitivities of photoreceptors and lamina monopolar cells in the dragonfly, Hemicordulia tau. *Journal of Comparative Physiology A Sensory Neural and Behavioral Physiology* **169**, 663-669.

Yang EC, Osorio D (1991b) Spectral Sensitivities of Photoreceptors and Lamina Monopolar Cells in the Dragonfly, Hemicordulia-Tau. *Journal of Comparative Physiology a-Sensory Neural and Behavioral Physiology* **169**, 663-669.

Yang J, Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* **43**, W174-181.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591.

Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. *Mol Biol Evol* **28**, 1217-1228.

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641-1650.

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908-917.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107-1118.

Yi H, Jin L (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res* **41**, e75.

Yokoyama S (2008) Evolution of dim-light and color vision pigments. *Annu Rev Genomics Hum Genet* **9**, 259-282.

Yuan F, Bernard GD, Le J, Briscoe AD (2010) Contrasting modes of evolution of the visual pigments in Heliconius butterflies. *Mol Biol Evol* **27**, 2392-2405.

Zhang J, Kumar S, Nei M (1997) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol* **14**, 1335-1338.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**, 2472-2479.

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* **95**, 3708-3713.

Zhang JZ (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292-298.

Zhao QY, Wang Y, Kong YM*, et al.* (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12 Suppl 14**, S2.

Figures and Tables

Figures

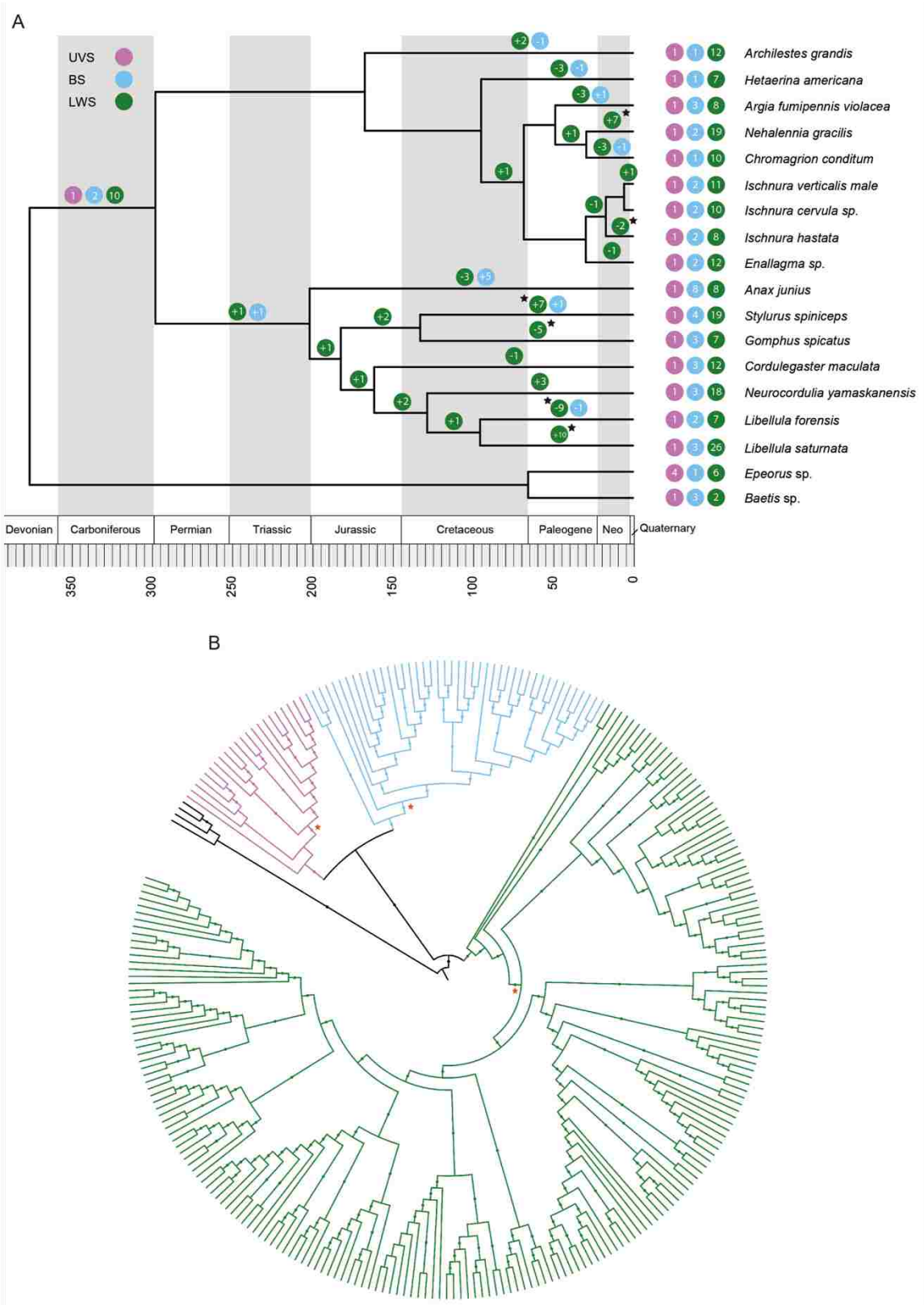


Fig. 1. Odonata opsin evolution (A) Opsin gains and losses along the dated Odonata species tree. All significant opsin turnovers occurred in terminal branches after 150 Mya. Significant opsin turnovers (Viterbi method with the randomly generated likelihood distribution, $P < 0.05$) are denoted by black stars. (B) Inferred ML opsin evolutionary relationships of Odonata (+ Ephemeroptera and mollusks as outgroups) with bootstrap >80% indicated by dots. Strong positive selection was present on UVS (purple), BS (blue) and LWS (green) ancestral branches as well as amino acid sites. Branches tested for positive selection using the PAML branch-site model A are identified by orange stars.

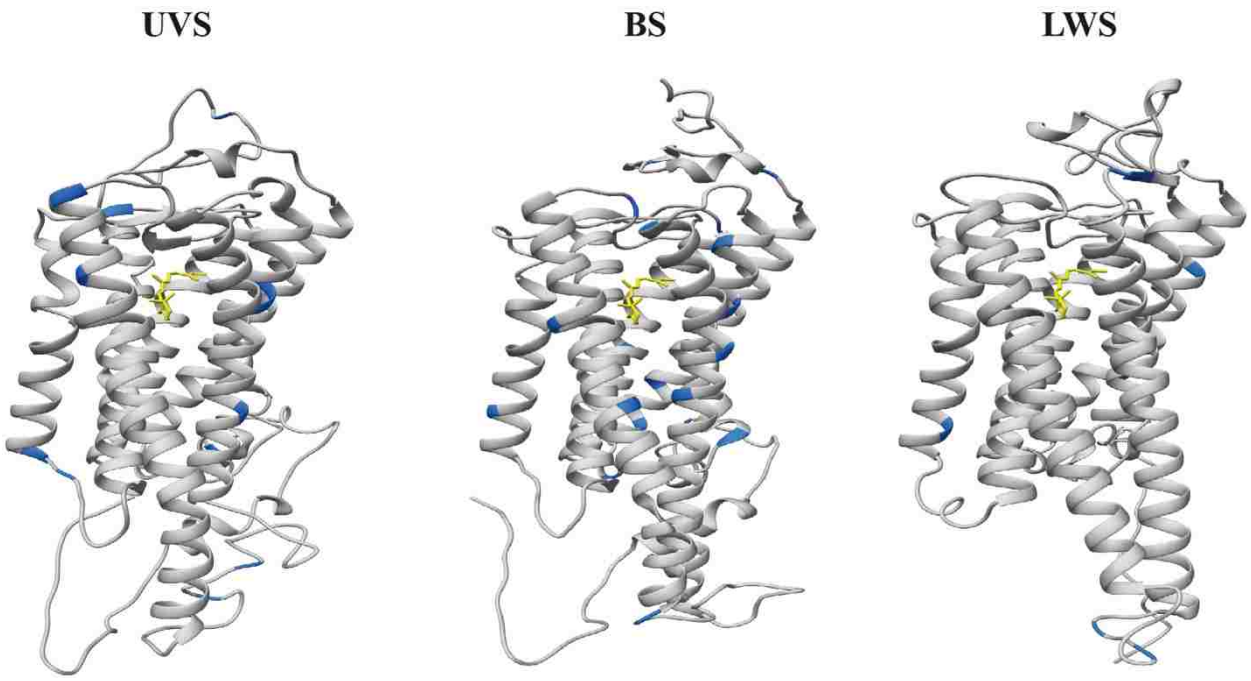


Fig. 2. Predicted 3D models of reconstructed ancestral opsin proteins with positively selected sites (dN/dS (ω) > 1, blue) indentified by BEB algorithm (P > 0.95). Yellow structure represents a chromophore molecule. No sites within the binding pocket were found evolving under positive selection.

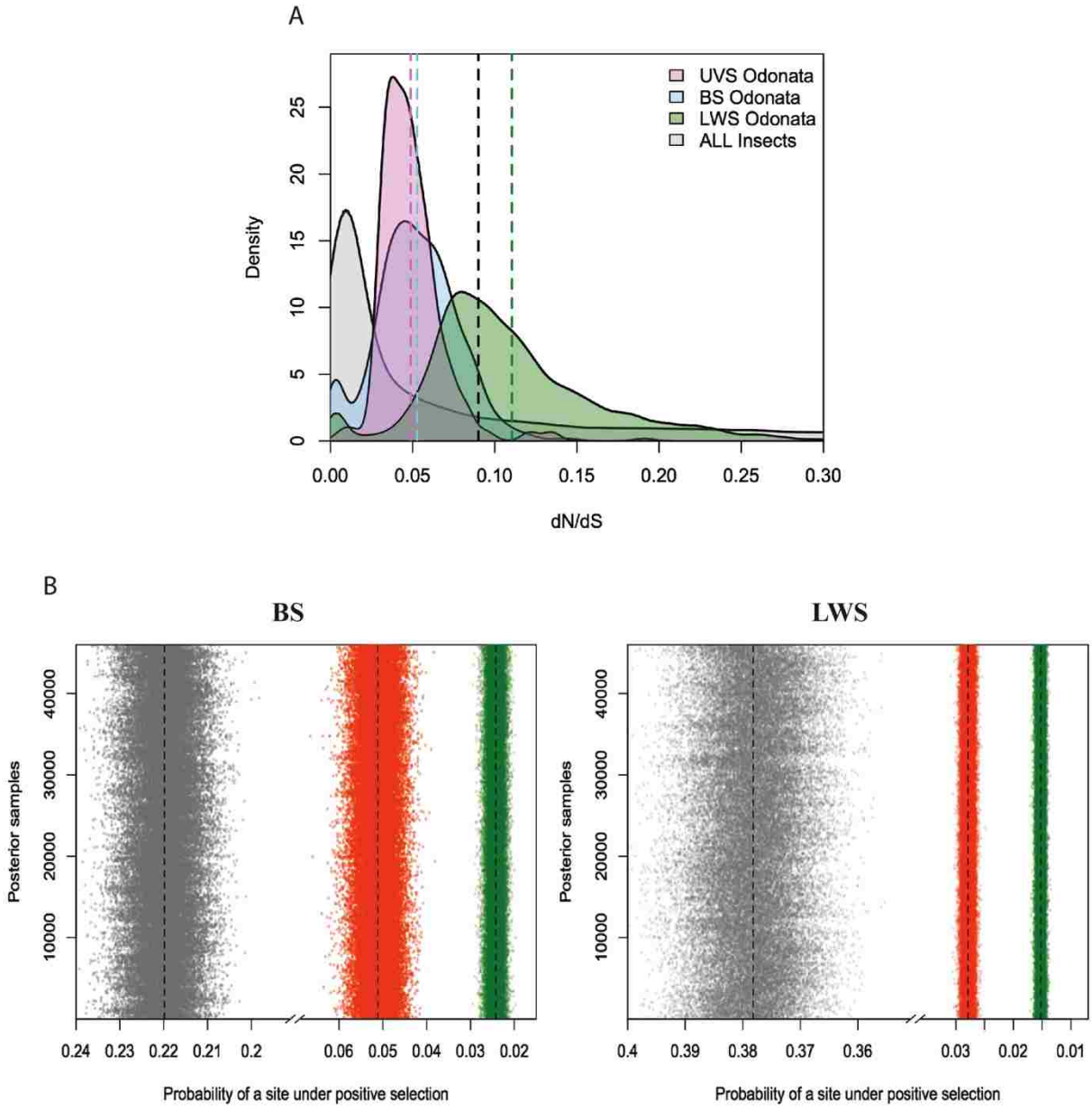


Fig. 3. (A) Distributions of pairwise dN/dS (ω) ratios calculated for each opsin class in Odonata and for all insect opsins. Dashed lines represent distribution means. Purifying selection is more relaxed in the LWS class than in the BS, UVS or all insect opsins. (B) Posterior samples of probabilities of a site that is under positive selection before (grey), after duplication events (red) and terminal branches (green) (note that x axis is inverted since intensity of positive selection "diffuses" toward the tips of the tree (i.e. from the left to the right)). The strength of positive

selection is quickly weakening towards the tips of the tree according to a diffusion model. (Note that the purpose of the "diffusion" model was to compare the magnitude of positive selection within opsin classes but not between).

Tables

Table 1. Categories of gene duplication models (Innan and Kondrashov 2010). Pre-duplication Phase (PDP), Positive Selection (PS) and Selection Pressure (SP)

Model Category	Molecular Evolution: Pre-duplication/Fixation Phase	Molecular Evolution: Preservation Phase		Functional Divergence
		Old Copy	New Copy	
Category I	Neutrality	same as PDP /can be accelerated by PS	can be accelerated by PS	neo/subfunctionalization
Category II	PS on duplication	same as PDP/ SP can be relaxed	SP can be relaxed / can be accelerated by PS	original/ neofunctionalization
Category III	pre-duplication PS on allelic variation	same as PDP/can be accelerated by PS/always accelerated by PS	can be accelerated by PS/always accelerated by PS	original/neo/sub/multi-functionalization
Category IV	NA	SP can be relaxed	SP can be relaxed	original

Table 2. Analyses of Positive Selection using Branch-Site Models (PAML) on UVS, BS and LWS Ancestral Branches of ML Opsin Gene Tree.

Branch[a]	Model	LnL	Site Class	Proportion (p)	Background (ω)	Foreground (ω)	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	A	-82682.800814	0	0.78274	0.07070	0.07070	**87**[b], 99, **198**, **288**, **291**, **339**, **340**, **393**	
			1	0.13481	1	1		
			2a	0.07034	0.07070	17.61057		
			2b	0.01211	1	17.61057		
	Null	-82687.018709	0	0.75702	0.07056	0.07056		
			1	0.13019	1	1		
			2a	0.09624	0.07056	1		
			2b	0.01655	1	1		
								0.00368
BS	A	-82671.822994	0	0.63906	0.07174	0.07174	**62**, **65**, **91**, 96, **102**, **119**, 135, 190, **228**, **244**, **254**, **277**, **318**, **323**, **336**, **349**, **355**, **370**, **390**	
			1	0.11022	1	1		
			2a	0.21384	0.07174	999[c]		
			2b	0.03688	1	999		
	Null	-82680.87305	0	0.60537	0.07133	0.07133		
			1	0.10424	1	1		
			2a	0.24774	0.07133	1		
			2b	0.04266	1	1		
								0.00002
UVS	A	-82681.344796	0	0.69447	0.07117	0.07117	**44**, 99, **142**, 183, 185, 232, **240**, **287**, 291, 316, **354**, **355**, **385**	
			1	0.11947	1	1		
			2a	0.15875	0.07117	4.66326		
			2b	0.02731	1	4.66326		
	Null	-82684.520754	0	0.65393	0.07109	0.07109		
			1	0.11253	1	1		
			2a	0.19926	0.07109	1		
			2b	0.03429	1	1		
								0.01173

LnL, ln likelihood; NA, not applicable; LRT, likelihood ratio test; BEB, Bayes Empirical Bayes

[a]Branches are marked in Fig. 2B and Fig. S2 (Supporting information)

[b]Positively selected sites (highlighted in bold) that agree with the inferred positively selected sites using all opsin sequences plus opsin sequences from (Futahashi *et al.* 2015) dataset (Alignment 3)

[c]Values of 999 for ω indicate dS = 0, so ω is undefined.

Table 3. Analyses of Positive Selection using Random Site Models (PAML) on UVS, BS and LWS ML opsin trees.

Opsin Class	Site model	LnL	Site class	Proportion (p)	Ω	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	M0	-46768.26547	0	NA	0.09621	NA	
	M1a	-44999.54374	0	0.84013	0.05610	NA	
			1	0.15987	1		
	M2a	-44985.58845	0	0.86001	0.05623	366, 376	4.3×10^{-8} (M2 vs. M1a)
			1	0.13999	1		
			2	0	57.25390		
	M3	-44022.62930	0	0.57428	0.01165	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.30869	0.15666		
			2	0.11703	0.70782		
	M7	-43848.98755	0-9	0-9:0.1	0 0.00015 0.00167 0.00826 0.02719 0.07019 0.15379 0.29834 0.52413 0.83686	NA	
	M8	-43813.60528	0-10	0-9:0.097, 10:0.02997	0.00001 0.00034 0.00244 0.00891 0.02360 0.05185 0.10133 0.18430 0.32449 0.59364 1.30216	3, 21, 366, 368, 376	$< 10^{-5}$ (M8 vs. M7)
BS	M0	-17851.24946	0	NA	0.08251	NA	
	M1a	-17586.53029	0	0.88206	0.0653	NA	
			1	0.11794	1		
	M2a	-17586.53029	0	0.88206	0.0653		> 0.99 (M2 vs. M1a)
			1	0.03149	1		
			2	0.08645	1		
	M3	-17228.57548	0	0.46133	0.00358	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.37380	0.09126		
			2	0.16486	0.36698		
	M7	-17220.93298	0-9	0-9:0.1	0.00002 0.0007 0.00362 0.01075 0.02451 0.04805 0.08606 0.14678 0.24851 0.45904	NA	
	M8	-17220.77872	0-10	0-9: 0.09962,10: 0.00384	0.00002 0.00074 0.00371	NA	0.579

					0.01079 0.02420 0.04688 0.08322 0.14099 0.23777 0.43986 1		(M8 vs. M7)
UVS	M0	-6991.807219	0	NA	0.06167	NA	
	M1a	-6913.55876	0	0.93414	0.03948	NA	
			1	0.06586	1		
	M2a	-6913.55876	0	0.93415	0.03948		> 0.99 (M2 vs. M1a)
			1	0.00039	1		
			2	0.06546	1		
	M3	-6836.192312	0	0.70175	0.00269	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.26312	0.17383		
			2	0.03513	0.59783		
	M7	-6838.88333	0-9	0-9:0.1	0 0.00001 0.00011 0.00076 0.00335 0.01098 0.02986 0.07219 0.16517 0.40289	NA	
	M8	-6838.258052	0-10	0-9:0.09941, 10:0.00591	0 0.00001 0.00013 0.00086 0.00353 0.01098 0.02860 0.06688 0.14947 0.36258 1	NA	0.263 (M8 vs. M7)

LnL, ln likelihood; NA, not applicable; LRT, likelihood ratio test; BEB, Bayes Empirical Bayes

Table 4. Category III evolutionary models of gene duplication that lack fate-determination phase (Innan and Kondrashov 2010). Pre-duplication Phase (PDP) and Positive Selection (PS).

Gene Duplication Model	Selection pressure on pre-duplicational variation	Selection pressure on a new copy	Selection pressure on the old copy
Permanent heterozygote	PS	can be accelerated by PS	can be accelerated by PS
Adaptive radiation	PS	same as PDP	can be accelerated by PS
Multiallelic diversifying selection	PS	always accelerated by PS	always accelerated by PS

Appendix Materials and Methods and Results/Discussion

Materials and Methods

ORF annotation

The subset of the longest ORFs is utilized to empirically estimate parameters for a Markov model based on hexamer distribution. The reference null distribution that represents non-coding sequences is constructed by randomizing the composition of these longest contigs. During the next decision step, each longest determined ORF and its 5 other alternative reading frames are tested using the trained Markov model. If the log-likelihood coding/noncoding ratio is positive and is the highest, this putative ORF with the correct reading frame is retained in the protein collection (proteome).

Orthology Assignment and Cluster Filtering

Only TransDecoder-predicted protein sequences were considered to represent our conserved data set for further orthology detection, phylogenetic inference and evolutionary analyses. In parallel, an alignment-free approach was taken using untrimmed Trinity transcriptome assemblies (see Phylogenetic Analyses section). To group protein sequences into gene families across multiple taxa, we exploited performance of several broadly used heuristic best-match orthology prediction tools (Kristensen *et al.* 2011), namely InParanoid-MultiParanoid v4.1 (Alexeyenko *et al.* 2006; Remm *et al.* 2001), OrthoMCL (Li *et al.* 2003) and HaMStR v13.2.2 (Ebersberger *et al.* 2009). Here we distinguish three fundamental types of inferred gene families such as 1:1 orthology (1:1) where each taxon is represented exactly once (single-copy), 1:1 with gene loss/missing data orthology (1:1-LM) and paralogy where at least one taxon is represented two or more times. Since homologous relationships between genes include

orthologous as well as paralogous interactions we collectively call all the above defined clusters as homology clusters.

All amino acid sequences of putative homologous genes inferred by InParanoid, OrthoMCL or HaMStR were individually aligned to form multiple sequence alignment (MSA) homology clusters for the subsequent analyses using MAFFT v. 6.864b (Katoh *et al.* 2002) with the "-auto" flag that enabled detection of the best alignment strategy between accuracy- and speed-oriented methods. Then, all the clusters were further filtered to reduce false positive homology assignments (Fujimoto *et al.* 2016a) using machine learning logistic regression classifier of OGCleaner (Fujimoto *et al.* 2016b)

InParanoid-MultiParanoid

InParanoid (Ostlund *et al.* 2010) initially performs bidirectional BLASTP between two proteomes to detect bidirectional BLASTP hits in the pairwise manner. For this step, we set default parameters with the BLOSUM62 protein substitution matrix and bit score cutoff of 40 for all-against-all BLASTP searches. Next, MultiParanoid forms multi-species groups using the notion of a single-linkage. Due to inefficient MultiParanoid clustering algorithm, we had to perform a transitive closure to compile homology clusters for all species together. Transitive closure is an operation performed on a set of related values. Formally, a set S is transitive if the following condition is true: for all values A, B, and C in S, if A is related to B and B is related to C, then A is related to C. Transitive closure takes a set (transitive or non-transitive) and creates all transitive relationships, if they do not already exist. When a set is already transitive, its transitive closure is identical to itself. In the case of the pairwise relationships produced by InParanoid, we constructed orthologous clusters using the notion of transitive closure, where gene identifiers were the values, and homology was the relationship. Our data set consisted of N

= 20 proteomes (18 Odonata + 2 Ephemeroptera), so we had to perform N × (N - 1)/2 = 190 pairwise InParanoid queries.

OrthoMCL

OrthoMCL v2.0 (Li *et al.* 2003) was used to compute orthologous genes in all 20 species using predicted protein sequences by TransDecoder. In summary, these predicted sequences were used in an all-vs-all BLASTP to find putative orthologs and paralogs. Alignments were only considered with a E-value of 10^{-5} and lower. These sequences were then inserted into a graph that was resolved using the Markov Cluster algorithm. The Markov Cluster algorithm resolves the many-to-many orthologous relationships in the graph by simulating random walks on the graph resulting in clusters of proteins. The output clusters were split into two data sets: original clusters and clusters with removed Trinity-predicted isoform sequences (see Trinity manual for details) retaining the longest one. Doing so we increased number of 1:1 clusters originated from false paralogy clusters.

HaMStR

To delineate putative orthologous sequences in the proteome sets, 5332 core 1:1 ortholog groups of 5 arthropod primer species (*Ixodes scapularis, Daphnia pulex, Rhodnius prolixus, Apis mellifera* and *Heliconius melpomene*) for training profile hidden Markov Models (pHMM) were retrieved from the OrthoDB v7 (Waterhouse *et al.* 2013). We used *Rhodnius prolixus* (triatomid bug) as the reference core proteome because this is the closest phylogenetically related species and publically available proteome to the Ephemeroptera/Odonata lineage (Meusemann *et al.* 2010). Each core orthology cluster was aligned to create MSA using MAFFT and converted into HMM profile using HMMER v. 3.0 (Eddy 2011). Bidirectional BLASTP hits against the reference proteome were derived using reciprocal BLASTP with the default parameters. Note

however that HaMStR (Ebersberger *et al.* 2009) along with the most representative best-hit orthologs (flag 1) retains non-representative co-orthologs (flag 0), which are usually excluded from the analysis, although it might be assumed that co-orthologs can exhibit significant true homologous interactions with the representative ortholog.

Thus we created two different types of clusters: conserved 1:1 and 1:1-LM clusters where no co-orthologs were mapped and relaxed 1:1 and 1:1-LM clusters were formed by excluding all non-representative co-orthologs and keeping exclusively representative orthologs.

Cluster filtering using machine learning approach

We filtered the putative orthology clusters by using a machine learning technique previously developed in (Fujimoto *et al.* 2016a). This method varies from heuristic based approaches by training a machine learning algorithm that is then able to differentiate between true homology and false homology clusters. The clusters of peptide sequences found in OrthoDB (Waterhouse *et al.* 2013) serve as positive examples of homology clusters. Selection of clusters from OrthoDB was limited to the entire arthropod phylogeny. Examples of non-homology clusters were generated by randomly drawing from all possible peptide sequences. Multiple sequence alignments (MSAs) were then calculated for each of the homology and non-homology clusters. With examples of both homology and non-homology clusters coupled with their respective MSAs, different attributes for each clusters were calculated and used as the input features for the learner. The machine learning algorithm was then trained using the OrthoDB homology clusters and the randomly generated non-homology clusters in order to classify novel instances as homology or non-homology clusters. Logistic regression served as the learning algorithm in this process.

Phylogenetic Analyses

For supermatrix (total evidence alignment) analyses, we removed ambiguously (randomly) aligned regions from all individual 1:1 and 1:1-LM gene alignments utilizing the trimming procedure of ALISCORE (Misof & Misof 2009), software based on the principle of parametric Monte Carlo resampling within a sliding window. This approach is more objective and exhibits less conservative behavior than commonly used non-parametric approaches implemented in GBLOCKS (Castresana 2000; Kuck *et al.* 2010). Then we separately concatenated trimmed alignments for each of the InParanoid, OrthoMCL and HaMStR outputs.

ML: IQ-TREE

To estimate putative species trees from the concatenated alignments we used a maximum likelihood (ML) approach implemented in IQ-TREE v0.9.6 (Nguyen *et al.* 2015) allowing internal protein model selection and execution with a best-fitted model according to the Bayesian Information Criterion (BIC) criterion. Nodal support was calculated for each tree using an ultrafast bootstrap (UFBoot) approach, which is robust against substitution model misspecification (Minh et al. 2013) with 1000 iterations.

Bayesian: ExaBayes

Bayesian supermatrix analyses were performed using ExaBayes v1.4.1 (Aberer *et al.* 2014). Our concatenated amino acid alignments were passed through ExaBayes under the default set of parameters with three independent Markov Chain Monte Carlo (MCMC) runs for 5×10^6 number of generations with sampling frequency of every 1000 generation. However, the tree search was set to start from a purely random topology rather than from a random-order addition parsimony tree to prevent any bias that might be caused by an informative starting tree. Then the

resultant trees from three independent runs were summarized into a majority rule consensus tree with drawn nodal posterior probabilities discarding 10% as burn-in.

Coalescent: ASTRAL with the multilocus bootstrapping

Coalescent-based species tree estimation was carried out in ASTRAL (Mirarab *et al.* 2014) with the multilocus bootstrapping method (Seo 2008) to draw nodal support. We applied ASTRAL algorithm individually to all 1:1 and 1:1+1:1-LM clusters of InParanoid, OrthoMCL or HaMStR outputs.

ASTRAL (Mirarab *et al.* 2014) is a statistically consistent summary method under multi-species coalescent model that also has been shown to be drastically more efficient and produce accurate results than other similar approaches on simulated conditions. Moreover, ASTRAL has an advantage of being a more accurate species tree estimator than supermatrix (concatenation) approach under at least moderate levels of ILS.

For multilocus bootstrapping, individual gene trees were estimated using IQ-TREE (Nguyen *et al.* 2015) forming input for ASTRAL. Each ML search was initiated with a selected best-fit substitution model with 200 standard bootstrap replications keeping the corresponding bootsrap trees (iqtree -s [1:1_cluster_alignment] -b 200 -m TEST). Second, 200 tree bootstrap files were organized by randomly associating 200 trees from the totality of all bootstrap trees available for 1:1 or 1:1+1:1-LM clusters. Then, ASTRAL mapped bootstrap probabilities on the "main" species tree estimated from 1:1 or 1:1+1:1-LM ML trees using the 200 bootstrap replicates (java -jar astral.4.7.6.jar -i [1:1_ML_trees] -b [bootstrap_trees] -r 200).

AF: Co-phylog

Additional to standard phylogenetic inferential approaches we applied an alignment-free (AF) species tree estimation algorithm using Co-phylog (Yi & Jin 2013). Co-phylog phylogenies were generated both from assembled DNA Trinity contigs and CDSs.

To find an optimal k-mer size at which phylogenetic signal is maximized, distinct k-mer counts were computed for each of the 20 taxa for k-mers ranging between 8 and 35. For both contigs and CDSs, distinct k-mer counts were maximized for k-mers ranging from 17 to 19. We thus selected 9 as the half context length k value required for Co-phylog. From the Trinity contigs a phylogenetic tree was created using 100 replicates generated for each sequence file and each replicate was a random sampling with replacement of sequences. For each set of replicates (a set containing one replicate from each species), Co-phylog was run using the same k-mer size as with the original tree. Thus 100 bootstrap trees were generated, one for each of the 100 replicate sets. These were used to compute bootstrap support values for nodes in the estimated tree. A Co-phylog phylogeny was created using the same method and k-mer size for CDSs.

Overall, for supermatrix analyses, concatenated individual gene alignments resulted in ten supermatrices of different sizes (Appendix Table 1), i.e. 1:1 and 1:1+1:1-LM for each of the orthology program outputs including additional clusters with isoforms excluded from OrthoMCL as well as relaxed clusters from HaMStR. Ten sets of ML gene trees inferred from above cluster data types were used to generate coalescent-based phylogenies. Thus, in total using contrasting methods (ML: IQ-TREE, Bayesian: ExaBayes, AF: Co-phylog and Coalescent: ASTRAL) 32 candidate species trees were estimated and compared to evaluate the robustness to variation in

orthology detection. For topological and tree length comparisons we used Robinson-Foulds (RF) (Robinson & Foulds 1981) distances and K scores (Soria-Carrasco *et al.* 2007), respectively.

Dating

Table 5. Fossil calibration points, age constraints and priors used for estimation of divergence times in MCMCTREE.

Fossil Taxon	Taxonomic Group	Assignment method	Age (MYA)	Prior	Reference
Saxonagrion minutus	Odonata	apomorphy	272.5	*SN(2.725, 0.5, 10)	(Nel *et al.* 1999)
Juragomphidae karatauensis	Anisoptera	apomorphy	155.7	SN(1.508, 0.3, 10)	(Nel *et al.* 2001)
Parahemiphlebia allendaviesi	Zygoptera	apomorphy	140.2	SN(1.402, 0.3, 10)	(Jarzembowski *et al.* 1998)
Epallagites avus	Calopterygidae	apomorphy	46.2	SN(0.462, 0.2, 10)	(Cockerell 1924)
Ischnura velteni	Ischnura	apomorphy	13.65	SN(0.1365, 0.2, 10)	(Bechly 2000)
Argia aliena	Argia + Nehalennia + Chromagrion	apomorphy	33.9	SN(0.339, 0.2, 10)	(Schwarz 1901)
Gomphus biconvexus	Gomphidae	apomorphy	125.45	SN(1.2545, 0.2, 10)	http://fossilworks.org
Cordulegaster intermedius	Cordulegastridae	apomorphy	145.5	SN(1.455, 0.2, 10)	(Nel *et al.* 1998)
Palaeolibellula zherikhini	Libellulidae	apomorphy	89.3	SN(0.893, 0.2, 10)	(Fleck *et al.* 1999)
Baetis gigantea	Baetis	apomorphy	33.9	SN(0.339, 0.2, 10)	
NA (95% confidence intervals)	Palaeoptera	NA	362 - 390	**B(3.62, 3.9)	(Misof *et al.* 2014; Tong *et al.* 2015)

*Skewed normal distribution, SN(location, scale, shape)

**Uniform distribution with soft upper and lower bounds (Yang & Rannala 2006)

Results/Discussion

Phylogenetic Inference

All approaches to phylogenetic reconstruction returned putative species trees that were concordant and generally topologically consistent with nearly identical branch lengths, uniformly high bootstrap supports and posterior probabilities (Appendix Figs. S1-S2 and Appendix S2). However tree estimates from relaxed HaMStR supermatrices usually supported alternative phylogenetic relationships indicated by non-zero normalized Robinson-Foulds (RF) distances (Appendix Fig. S1). The inaccurate detection of orthologous clusters using HaMStR is possibly due to several factors: 1) the relaxed criteria used for grouping HaMStR hits may result in clusters "contaminated" by random paralogous sequences (i.e., false positive 1:1 orthology) and/or 2) statistical inconsistency of unpartitioned supermatrix-driven ML/Bayesian phylogenetic reconstruction (Kolaczkowski & Thornton 2009; Roch & Steel 2014; Warnow 2012). The statistical consistency of partitioned data sets has not been verified within ML (Roch & Steel 2014), hence we did not apply any partitioning scheme to the supermatrices.

Since our taxon represents a highly diverged species from an ancient insect lineage, highly variable sites may reduce accuracy of phylogenetic inference. Nevertheless, the effect of such sites can be minimized by implementing of alignment-free (AF) methods that exhibit improved accuracy in the presence of high among-site variation (Hohl & Ragan 2007) especially for medium to large phylogenetic distances. Using AF algorithm of Co-phylog (Yi & Jin 2013),

we recovered species trees that exhibited full topological agreement with the majority of ML as well as Bayesian phylogenetic estimates (Appendix Fig. S1).

Additionally, a coalescent-based summary method (ASTRAL; (Mirarab *et al.* 2014)) was used to find the species tree from individual ML gene trees. The appealing property of ASTRAL is that its algorithm is statistically consistent compared to other available methods, under a multi-species coalescent model (Warnow 2015). Estimated ASTRAL trees from orthologous clusters predicted by different programs all agreed on an identical topology (RF = 0) (Appendix Fig. S1).

Evaluation of Gene Tree - Species Tree Discord

Gene genealogies (gene trees) were reconstructed from 1:1 orthologous clusters using an ML approach and compared to the species trees to evaluate possible discord. Almost all inferred gene genealogies were found to be topologically incongruent with the species trees. The empirical distributions of normalized RF distances exhibited statistically similar means (Wilcoxon rank sum test, adjusted $P > 0.12$) between OrthoMCL, InParanoid and HaMStR but differ significantly from the HaMStR relaxed distribution (Wilcoxon rank sum test, adjusted $P < 10^{-10}$, Appendix Fig. S3). This incongruence between gene and species trees may indicate that genes have been subjected to Incomplete Lineage Sorting (ILS). Usually, higher rates of ancestral polymorphism that leads to ILS, one of the major factors for gene conflict, pose serious challenges to species phylogenetic inference (Degnan & Rosenberg 2009). For instance, concatenated genes that were under ILS may yield an incorrect species phylogeny with high confidence (i.e. an incorrect topology with high bootstrap supports). Alternatively, genes may lack the phylogenetic signal required to resolve ancient divergence among Odonata (Salichos & Rokas 2013) that occurred at least ~298 MYA (Fig. 1A). We also note that higher pairwise topological disagreement among individual ML gene trees contributes to greater differences

between estimated species trees (Degnan & Rosenberg 2009; Edwards 2009) (Appendix Fig. S4). Moreover, despite the fact that mostly short branches are susceptible to ILS (Maddison 1997), we observed multiple disagreements at deeper nodes by evaluating how many times gene trees recover monophyly of each of the two suborders (Anisoptera and Zygoptera). Strikingly, more than 50% and 45% of gene trees did not exhibit monophyletic relationships for Anisoptera and Zygoptera, respectively. Further, the discrepancy between the number of genes that recover monophyly for the group is statistically significant (Fisher exact test, $P < 0.01$, Appendix Table S2), with the exception of when HaMStR clusters were used. Moreover, less than 20% of the genes recover both suborders as monophyletic. These results likely demonstrate differences in evolutionary constraints between the two suborders throughout their evolutionary histories.

Compared to many other insect orders, species within Odonata tend to have smaller population sizes relative to their divergence time, which in turn could reduce the effect of severe ILS (Maddison 1997), thus the observed gene-species tree discord might arise from other biological scenarios, e.g. intra- and interspecific ancestral gene flow. Alternatively, ancient rapid radiations and following multiple mass extinction events (Benton *et al.* 1993) could lengthen deep branches, so that these branches appear "immune" to effects of ILS when in fact they are not (e.g. deep discord at the subordinal level). Large amounts of sequence data were required to confidently resolve the Odonata species tree presented herein (Fig. S3, Supporting information). All trees are summarized in Appendix S2.

Anisoptera and Zygoptera Relationships

An explicit examination of the higher-level, e.g. suborder, evolutionary relationships of the order Odonata has been rarely addressed within the same phylogenetic estimate (Bybee *et al.* 2008; Carle *et al.* 2008). Most phylogenetic research has primarily focused on refining

phylogenetic structure within each of the major two suborders Anisoptera (e.g., (Misof 2002)) and Zygoptera (e.g., (Dijkstra *et al.* 2014a)) using mostly molecular data. Saux *et al.* (2003) were the first to use molecular data to examine the monophyly of both Zygoptera and Anisoptera using a single molecular marker, 12S rRNA. They recovered Zygoptera as a paraphyletic group with *Lestes disjunctus* (Lestidae) sister to a monophyletic Anisoptera. This result corroborated the findings of (Trueman 1996) who also recovered a non-monophyletic Zygoptera and monophyletic Anisoptera based strictly on morphological data. The hypothesis that Zygoptera was non-monophyletic was surprising since morphologically these two suborders can be clearly differentiated by differences in wing shape/venation, body size and relative position of the eyes to each other. Although it is almost certain that Anisoptera and Zygoptera are monophylyetic the data presented herein provide a clear, statistically significant result for the monophyly of both orders. We resolved the split between Anisoptera and Zygoptera using an unprecedented number of loci (Appendix Table S1) and high confidence (Fig. S3, Supporting information), suggesting that this is the true phylogenetic relationship between these groups

Appendix References

Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol Biol Evol* **31**, 2553-2556.

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15.

Bechly G (2000) A new fossil damselfly species (Insecta: Odonata: Zygoptera: Coenagrionidae: Ischnurinae) from Dominican amber. *Stuttgarter Beitraege zur Naturkunde Serie B (Geologie und Palaeontologie)* **299**, 1-9.

Benton MJ, Palaeontological Association., Royal Society (Great Britain), Linnean Society of London. (1993) *The Fossil record 2*, 1st edn. Chapman & Hall, London ; New York.

Bybee SM, Ogden TH, Branham MA, Whiting MF (2008) Molecules, morphology and fossils: a comprehensive approach to odonate phylogeny and the evolution of the odonate wing. *Cladistics-the International Journal of the Willi Hennig Society* **24**, 477-514.

Carle FL, Kjer KM, May ML (2008) Evolution of Odonata, with special reference to Coenagrionoidea (Zygoptera). *Arthropod Systematics and Phylogeny* **66**, 37-44.

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

Cockerell TDA (1924) Fossil insects in the United States National Museum. *Proceedings of the United States National Museum* **64**, 15 pp.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* **24**, 332-340.

Dijkstra KDB, Kalkman VJ, Dow RA, Stokvis FR, Van Tol J (2014) Redefining the damselfly families: a comprehensive molecular phylogeny of Zygoptera (Odonata). *Systematic Entomology* **39**, 68-96.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**, 157.

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1-19.

Fleck G, Nel A, Martinez-Delclos X (1999) The oldest record of libellulid dragonflies from the Upper Cretaceous of Kazakhstan (Insecta: Odonata, Anisoptera). *Cretaceous Research* **20**, 655-658.

Fujimoto MS, Suvorov A, Jensen NO, Clement MJ, Bybee SM (2016a) Detecting false positive sequence homology: a machine learning approach. *BMC Bioinformatics* **17**, 101.

Fujimoto MS, Suvorov A, Jensen NO*, et al.* (2016b) The OGCleaner: filtering false-positive homology clusters. *Bioinformatics*.

Hohl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology* **56**, 206-221.

Jarzembowski EA, Martinez-Delclos X, Bechly G*, et al.* (1998) The Mesozoic non-calopterygoid Zygoptera: description of new genera and species from the Lower Cretaceous of England and Brazil and their phylogenetic significance (Odonata, Zygoptera, Coenagrionoidea, Hemiphlebioidea, Lestoidea). *Cretaceous Research* **19**, 403-444.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066.

Kolaczkowski B, Thornton JW (2009) Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS One* **4**, e7891.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. *Brief Bioinform* **12**, 379-391.

Kuck P, Meusemann K, Dambach J*, et al.* (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool* **7**, 10.

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology* **46**, 523-536.

Meusemann K, von Reumont BM, Simon S*, et al.* (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* **27**, 2451-2464.

Mirarab S, Reaz R, Bayzid MS*, et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-548.

Misof B (2002) Diversity of Anisoptera (Odonata): infering speciation processes from patterns of morphological diversity. *Zoology (Jena)* **105**, 355-365.

Misof B, Liu S, Meusemann K*, et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767.

Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* **58**, 21-34.

Nel A, Bechly G, Jarzembowski E, Martinez-Delclos X (1998) A revision of the fossil petalurid dragonflies (Insecta: Odonata: Anisoptera: Petalurida). *Paleontologia Lombarda* **10**, 1-68.

Nel A, Bechly G, Martinez-Delclos X, Fleck G (2001) A new family of Anisoptera from the Upper Jurassic of Karatau in Kazakhstan (Insecta: Odonata: Juragomphidae n. fam.). *Stuttgarter Beitraege zur Naturkunde Serie B (Geologie und Palaeontologie)* **314**, 1-9.

Nel A, Gand G, Fleck G*, et al.* (1999) Saxonagrion minutus nov gen. et sp., the oldest damselfly from the Upper Permian of France (Odonatoptera, Panodonata, Saxonagrionidae nov fam.). *Geobios* **32**, 883-888.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274.

Ostlund G, Schmitt T, Forslund K*, et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**, D196-203.

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052.

Robinson DF, Foulds LR (1981) Comparison of Phylogenetic Trees. *Math Biosci* **53**, 131-147.

Roch S, Steel M (2014) Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol* **100C**, 56-62.

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327-331.

Saux C, Simon C, Spicer GS (2003) Phylogeny of the dragonfly and damselfly order odonata as inferred by mitochondrial 12S ribosomal RNA sequences. *Ann Entomol Soc Am* **96**, 693-699.

Schwarz O (1901) Some insects of special interest from Florissant, Colorado, and other points in the tertiaries of Colorado and Utah. *Bulletin of the United States Geological Survey* **No. 93**, 25 pp.

Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol* **25**, 960-971.

Soria-Carrasco V, Talavera G, Igea J, Castresana J (2007) The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics* **23**, 2954-2956.

Tong KJ, Duchene S, Ho SY, Lo N (2015) INSECT PHYLOGENOMICS. Comment on "Phylogenomics resolves the timing and pattern of insect evolution". *Science* **349**, 487.

Trueman JWH (1996) A preliminary cladistic analysis of odonate wing venation. *Odonatologica* **25**, 59-72.

Warnow T (2012) Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Curr* **4**, RRN1308.

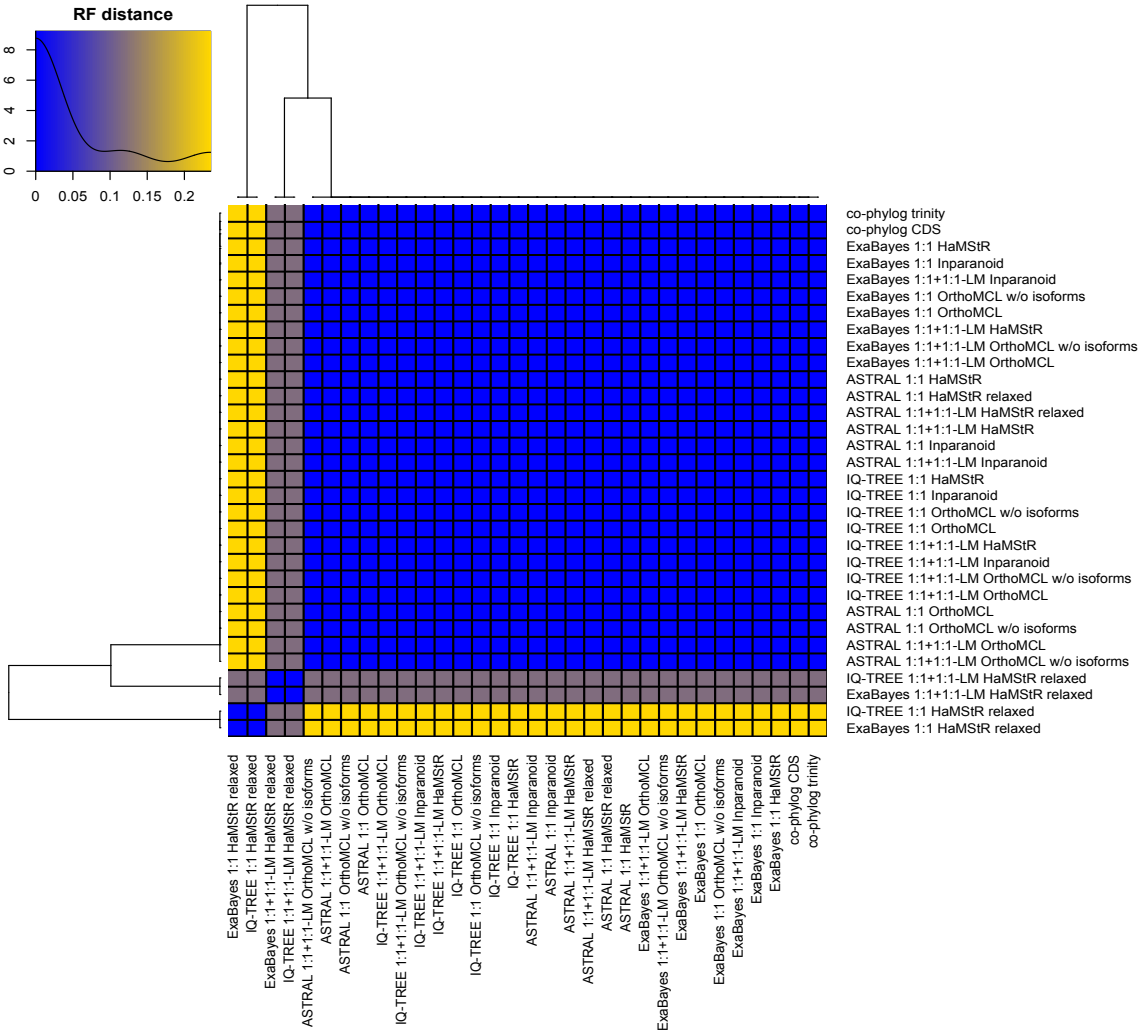
Warnow T (2015) Concatenation Analyses in the Presence of Incomplete Lineage Sorting. *PLoS Curr* **7**.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41**, D358-365.

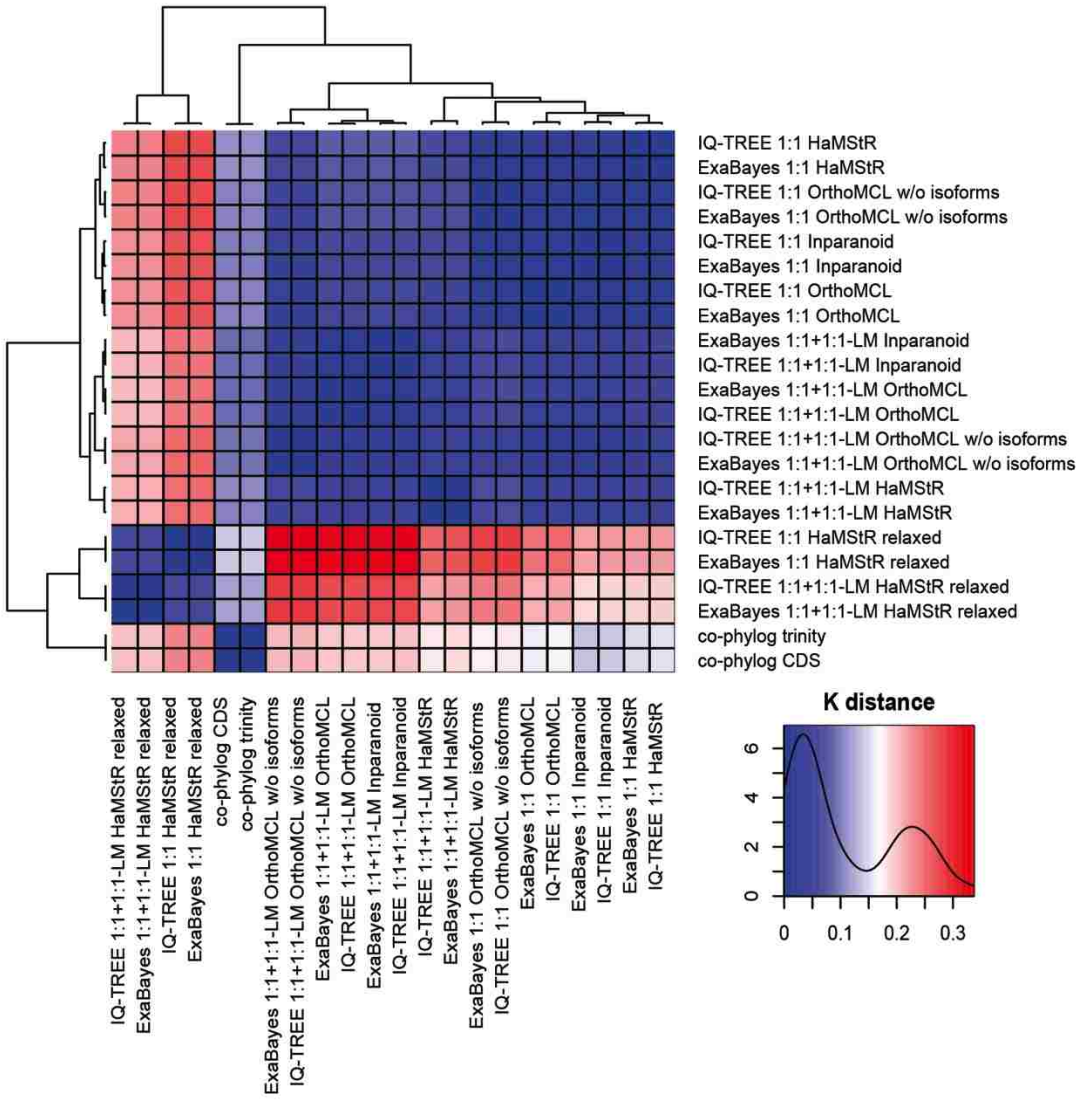
Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**, 212-226.

Yi H, Jin L (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res* **41**, e75.

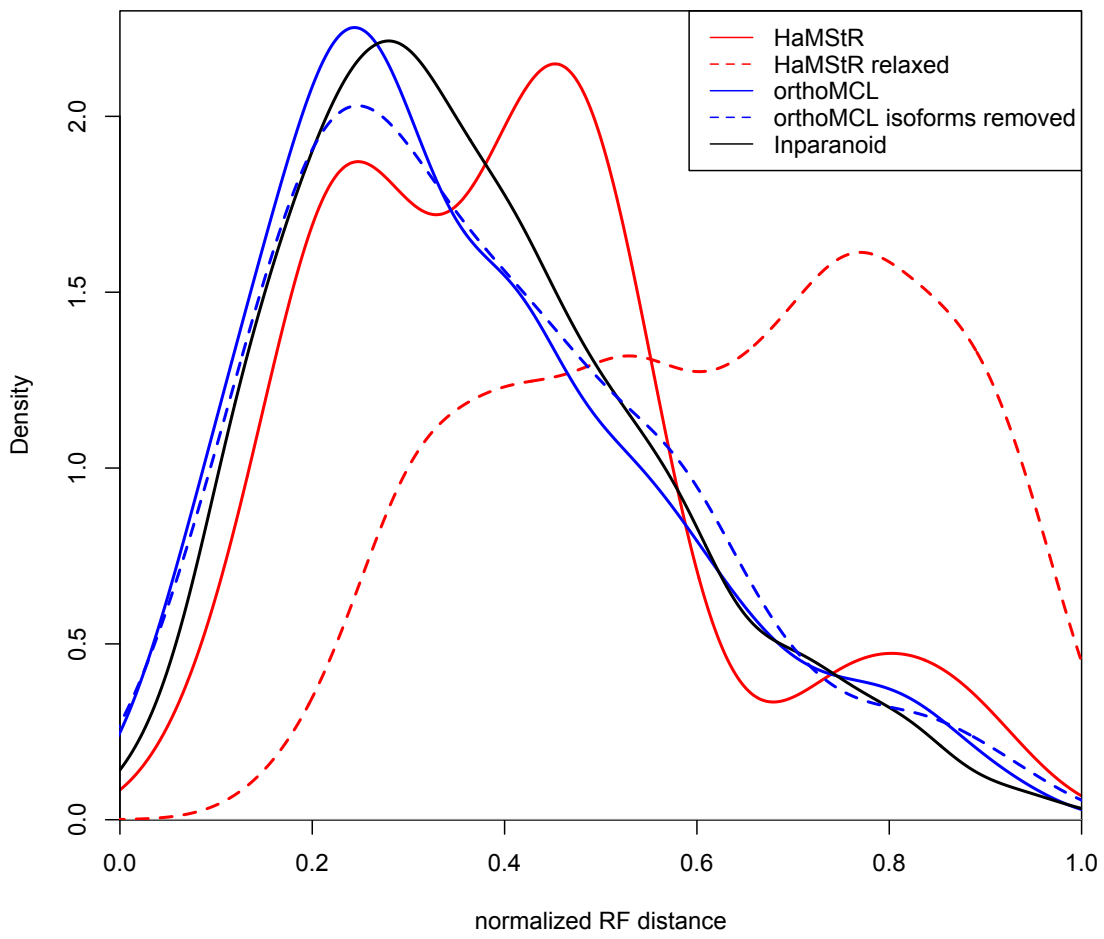
Appendix Tables and Figures



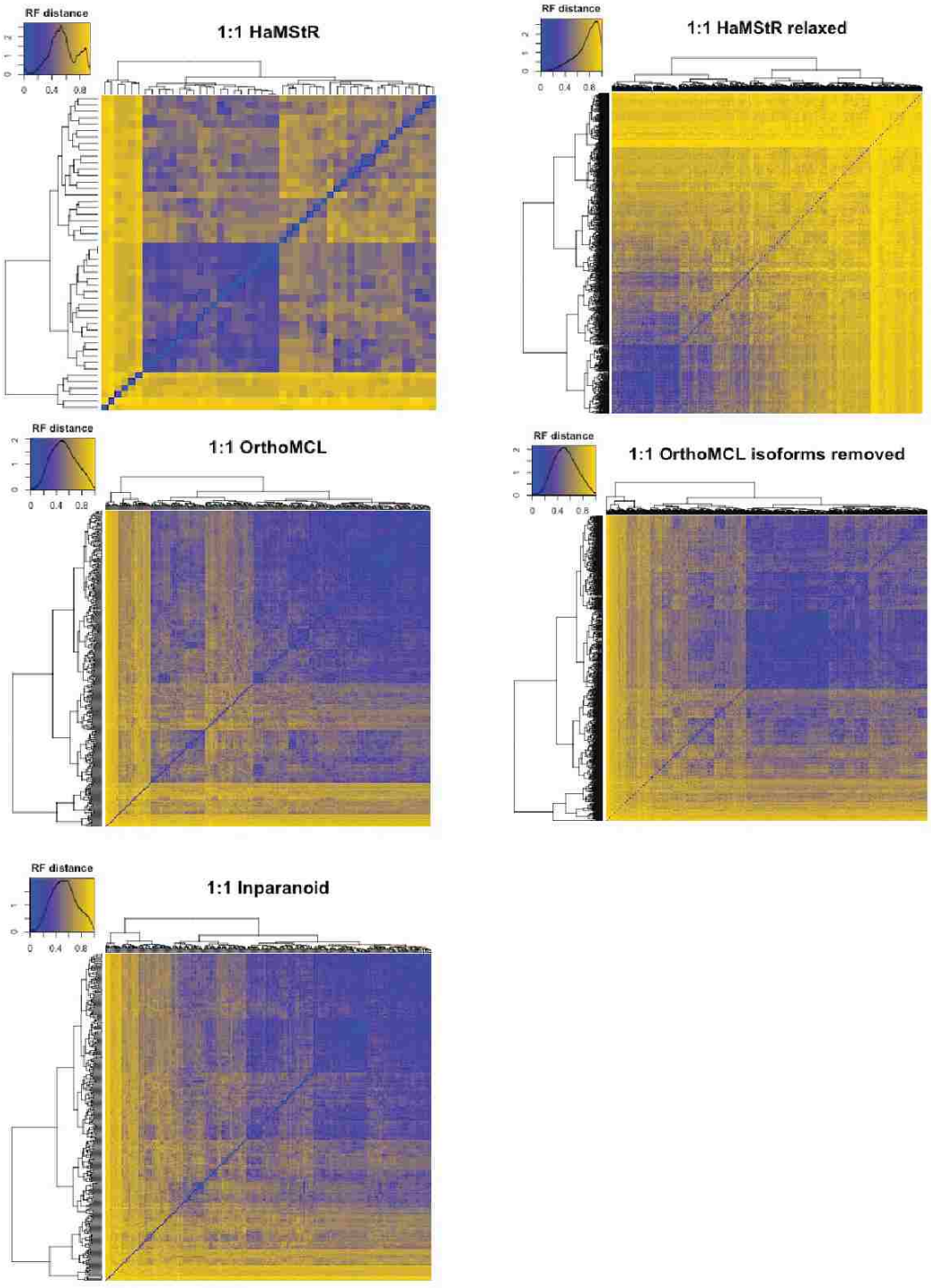
Appendix Figure S1. Hierarchical clustering of normalized Robinson-Foulds (RF) distances calculated from pairwise comparisons of 32 trees, estimated using different methodologies and data types. The upper corner box represents estimated kernel density of RF distances. Majority of the trees agree on one topology whereas small fraction of ML and Bayesian trees inferred from HaMStR relaxed clusters exhibit an alternative topology.



Appendix Figure S2. Hierarchical clustering of K scores (Soria-Carrasco *et al.* 2007) calculated between ML, Bayesian and Co-phylog species trees. Note that the K score metric is asymmetric and comparisons between tree1 and tree2 *versus* tree2 and tree1 might differ slightly. As determined from K score, phylogenetic inference from less stringent sequence data (HaMStR 1:1+1:1-LM relaxed datasets) is capable of providing a better approximation to a true species tree than more stringent data (HaMStR 1:1 relaxed datasets). Overall nearly all phylogenetic analyses returned congruent topologies with only minor among-branch variation.



Appendix Figure S3. Kernel density estimates of normalized RF distances generated by comparison of the ML species tree against all ML gene trees inferred from different cluster types.



Appendix Figure S4. Topological pairwise comparisons between ML individual gene trees.

Appendix Table S1. Cluster and supermatrix statistics and substitution models.

Supermatrix	Number of Gene Clusters	Number of Gene Clusters after Machine Learning Ttrimming	Number of Genes	Size (Peptide Bases)	Size after ALICUT	IQ-TREE Best-fit Substitution Model	IQ-TREE LogL	Number of Genes with 50% or more Taxa for ASTRAL
1:1 HaMStR	291	286	49	16215	15086	LG+I+G4	-111933.1629	49
1:1+1:1-LM HaMStR			286	82384	52090	JTTDCMut+G4 +F	-409241.8192	198
1:1 HaMStR relaxed	2121	1853	844	652029	243725	VT+G4+F	-3074595.527	844
1:1+1:1-LM HaMStR relaxed			1853	1170313	380995	VT+G4+F	-4440397.093	1667
1:1 Inparanoid	8957	8939	298	117925	104577	JTTDCMut+I+ G4+F	-821852.5294	298
1:1+1:1-LM Inparanoid			8939	2191104	1166243	JTT+G4+F	-6236486.942	1037
1:1 OrthoMCL	9953	9858	343	128581	110620	JTTDCMut+I+ G4+F	-919382.0808	343
1:1+1:1-LM OrthoMCL			9858	2578751	1187631	JTT+G4+F	-6702709.049	1465
1:1 OrthoMCL w/o isoforms	11723	11188	770	355605	285648	JTT+G4+F	-2432747.342	770
1:1+1:1-LM OrthoMCL w/o isoforms			11188	3071775	1589798	JTT+G4+F	-9853800.848	2471

Appendix Table S2. Monophyly tests of Anizoptera and Zygoptera lineages

Cluster Type	Monophyletic Anisoptera	Non-Monophyletic Anisoptera	Monophyletic Zygoptera	Non-Monophyletic Zygoptera	*FET P value (Anisoptera vs. Zygoptera)	Monophyletic Both
1:1 HaMStR	15 (0.3061224)	34 (0.6938776)	23 (0.4693878)	26 (0.5306122)	0.1463	5 (0.1020408)
1:1 HaMStR relaxed	183 (0.2168246)	661 (0.7831754)	236 (0.2796209)	608 (0.7203791)	0.003363	67 (0.07938389)
1:1 Inparanoid	120 (0.4026846)	178 (0.5973154)	172 (0.5771812)	126 (0.4228188)	2.81E-05	57 (0.1912752)
1:1 OrthoMCL	141 (0.4110787)	202 (0.5889213)	196 (0.5714286)	147 (0.4285714)	3.59E-05	64 (0.1865889)
1:1 OrthoMCL w/o isoforms	325 (0.4220779)	445 (0.5779221)	404 (0.5246753)	366 (0.4753247)	6.77E-05	123 (0.1597403)

* Fisher Exact Test, FET

Supplementary Figures and Tables

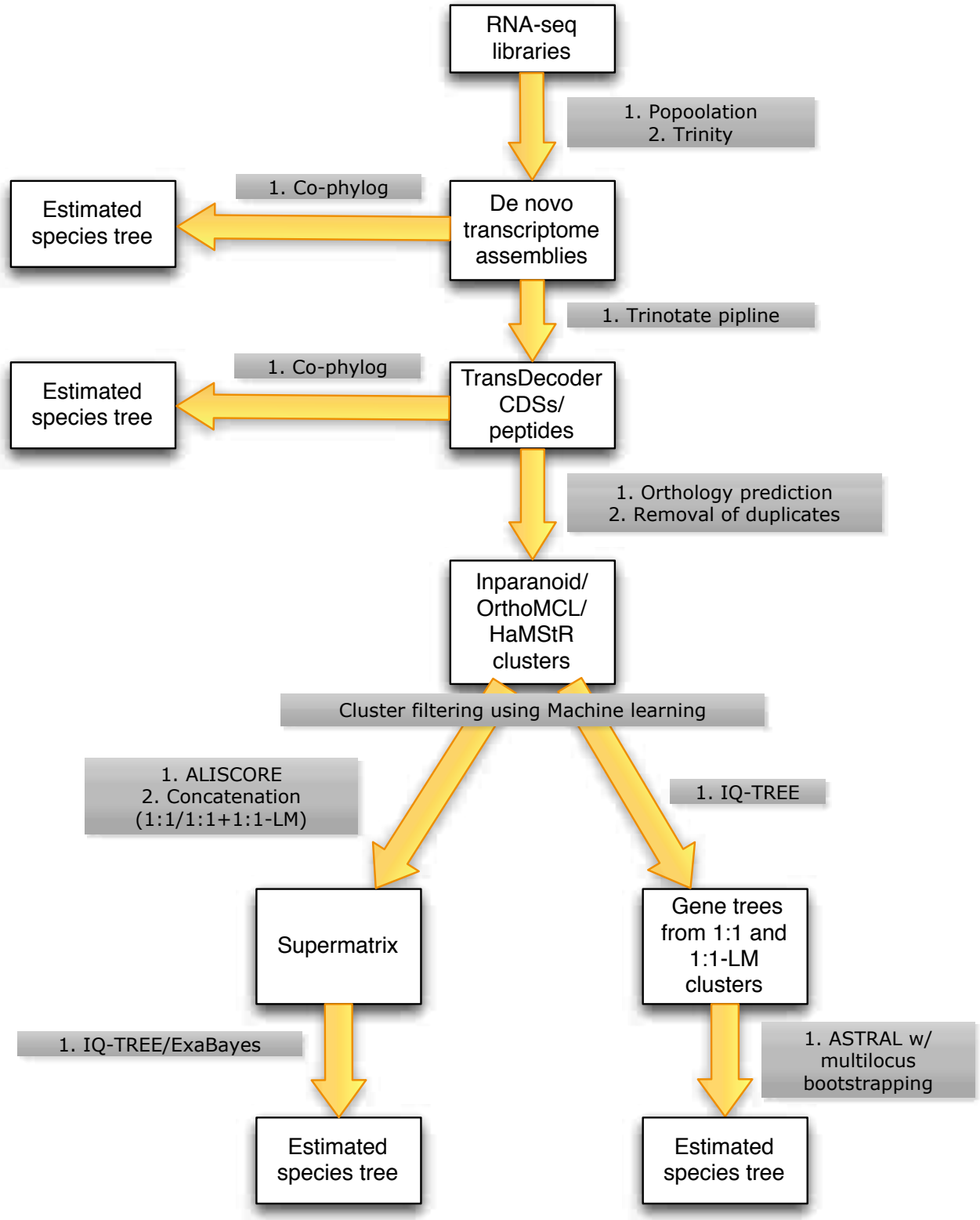


Figure S1. Pipeline of phylogenetic analyses.

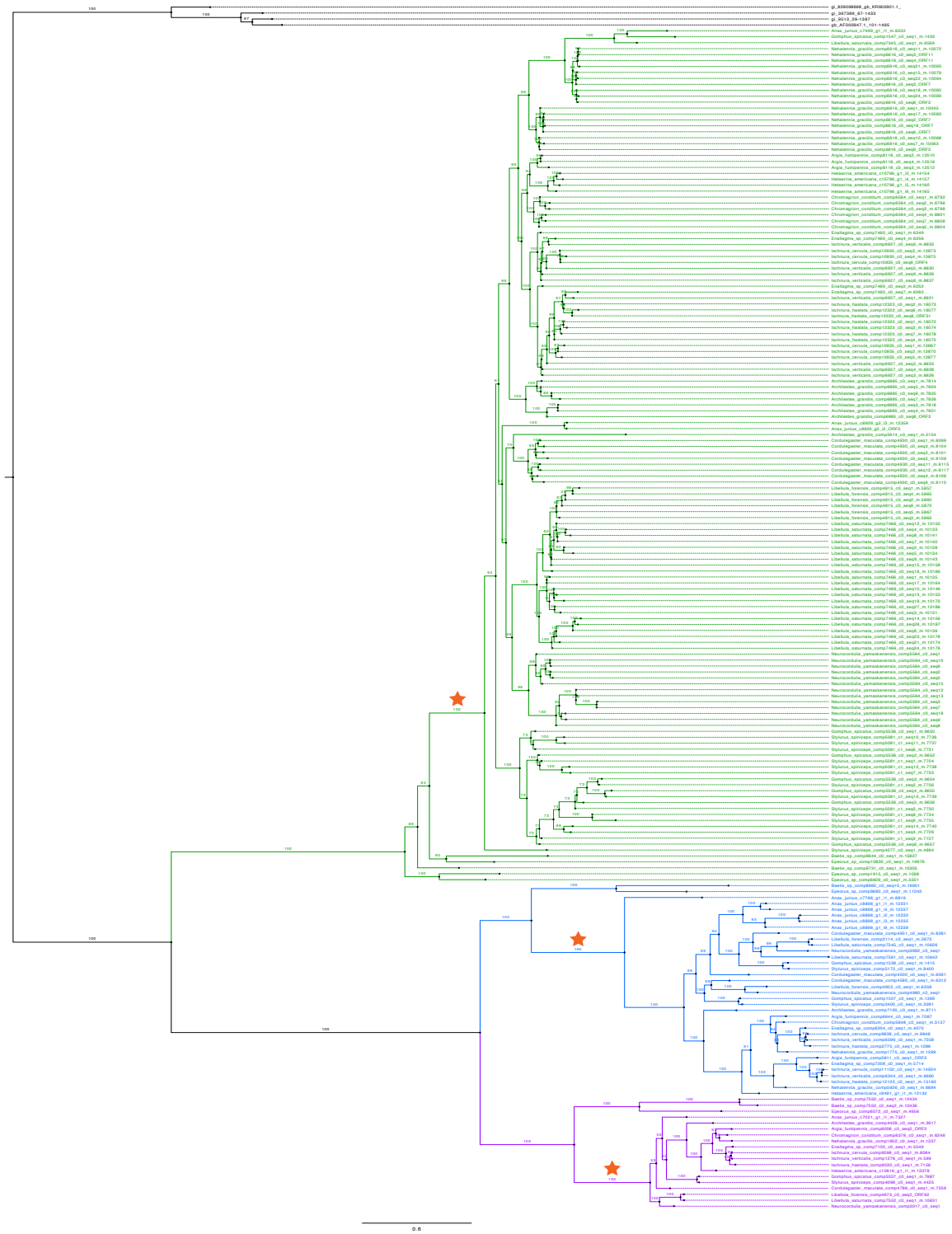


Figure S2. ML opsin tree. Opsin ancestral lineages are marked with the stars (see Alignment 1).

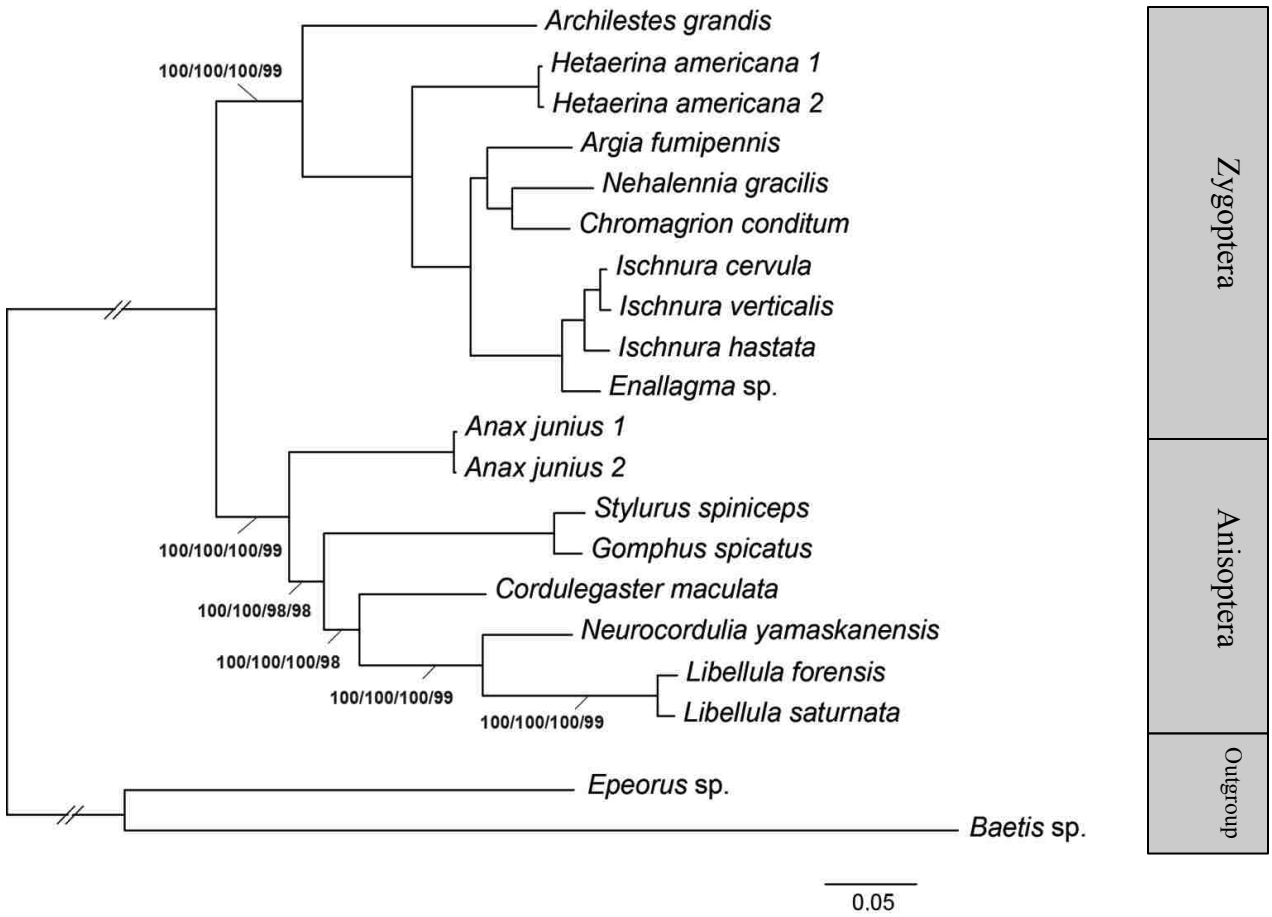


Figure S3. Maximum Likelihood (ML) phylogenetic tree inferred for Odonata (plus Ephemeroptera outgroup) using a concatenated protein supermatrix of 770, 1:1 OrthoMCL orthologous clusters with randomly aligned sites excluded. Phylogenetic supports are indicated as follows: IQ-TREE UFboot/ExaBayes posterior probability/multilocus ASTRAL bootstrap/Co-phylog bootstrap. Only <100 supports are shown.

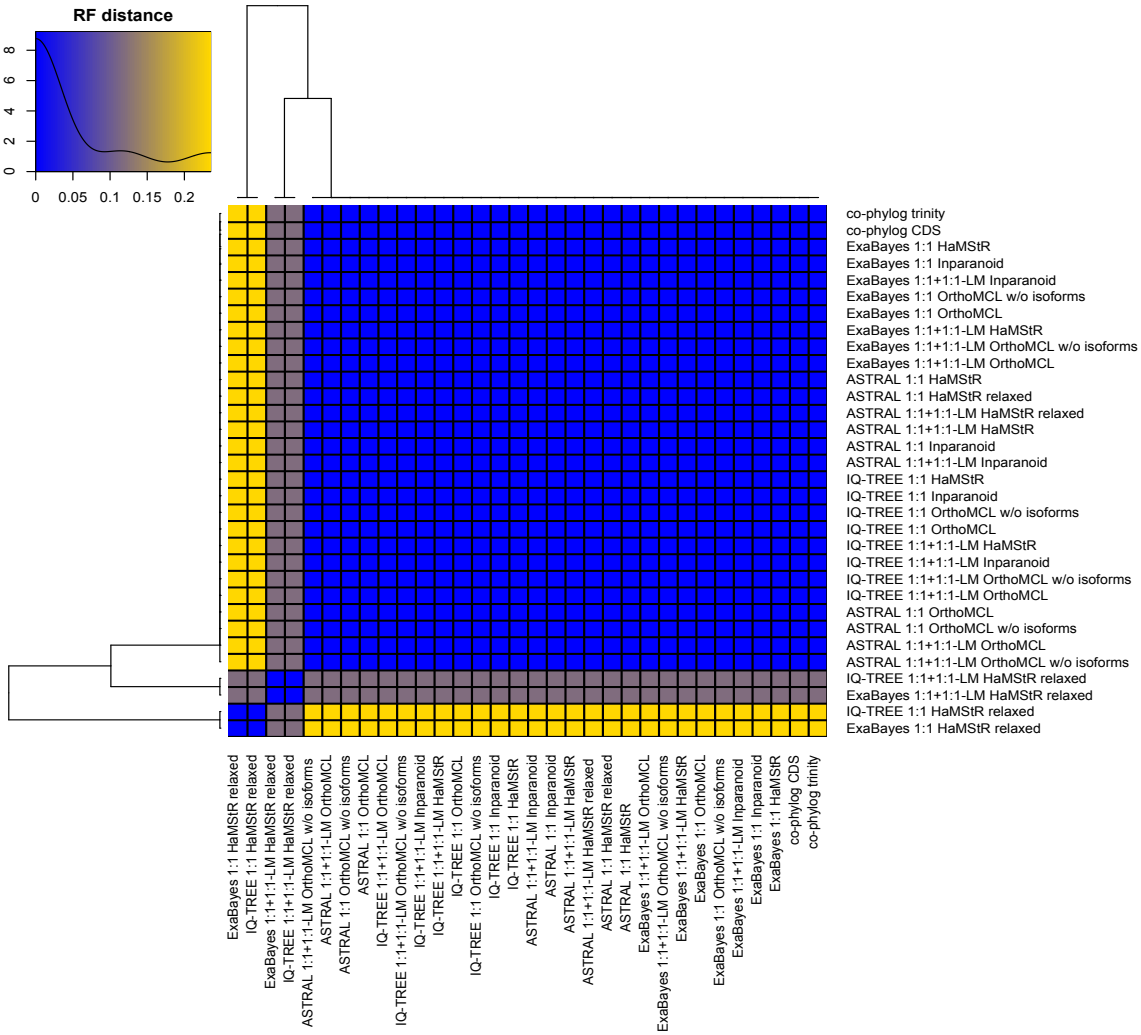


Figure S4. Hierarchical clustering of normalized Robinson-Foulds (RF) distances calculated from pairwise comparisons of 32 trees, estimated using different methodologies and data types. The upper corner box represents estimated kernel density of RF distances. Majority of the trees agree on one topology whereas small fraction of ML and Bayesian trees inferred from HaMStR relaxed clusters exhibit an alternative topology.

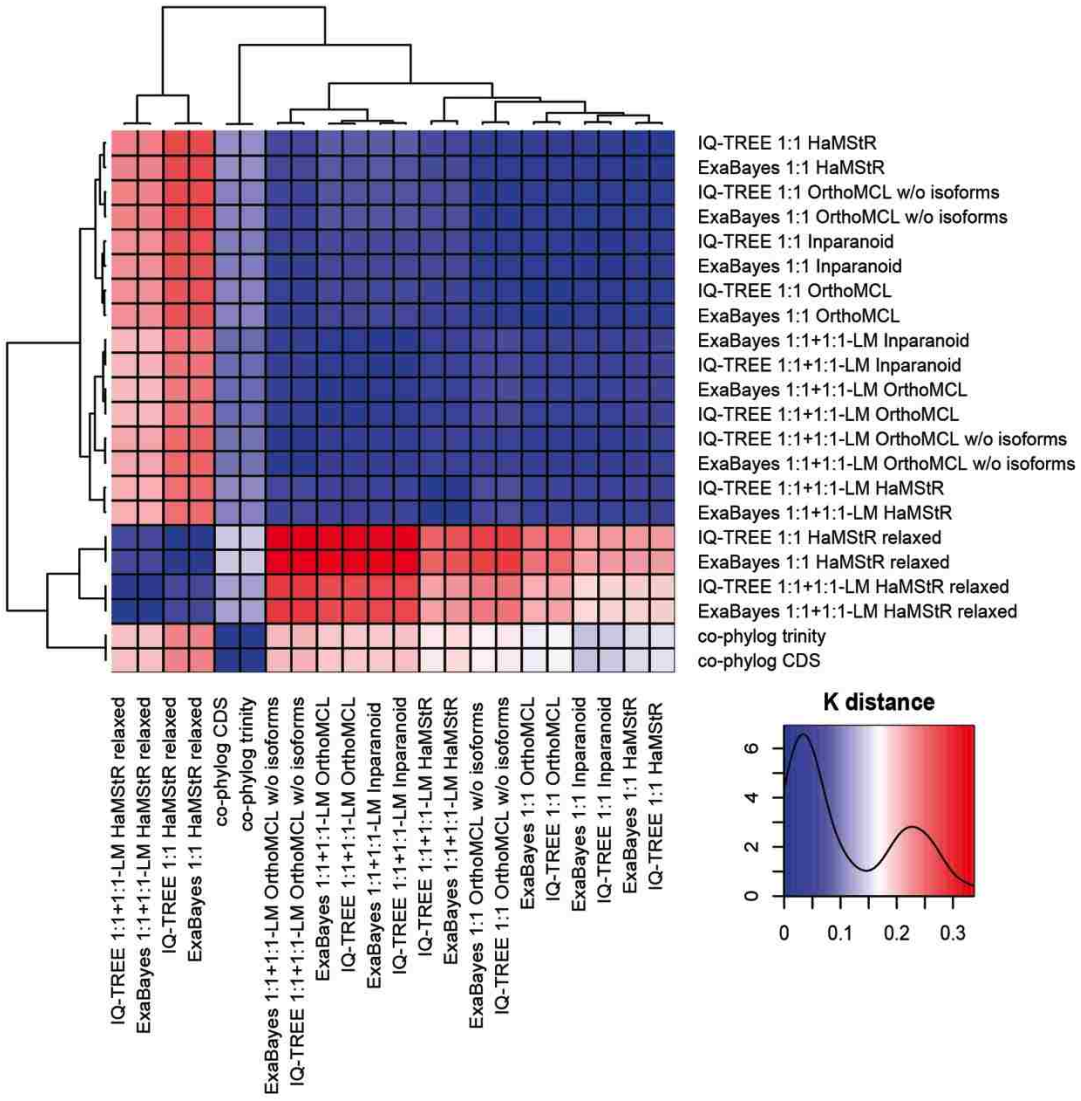


Figure S5. Hierarchical clustering of K scores (Soria-Carrasco *et al.* 2007) calculated between ML, Bayesian and Co-phylog species trees. Note that the K score metric is asymmetric and comparisons between tree1 and tree2 *versus* tree2 and tree1 might differ slightly. As determined from K score, phylogenetic inference from less stringent sequence data (HaMStR 1:1+1:1-LM relaxed datasets) is capable of providing a better approximation to a true species tree than more stringent data (HaMStR 1:1 relaxed datasets). Overall nearly all phylogenetic analyses returned congruent topologies with only minor among-branch variation.

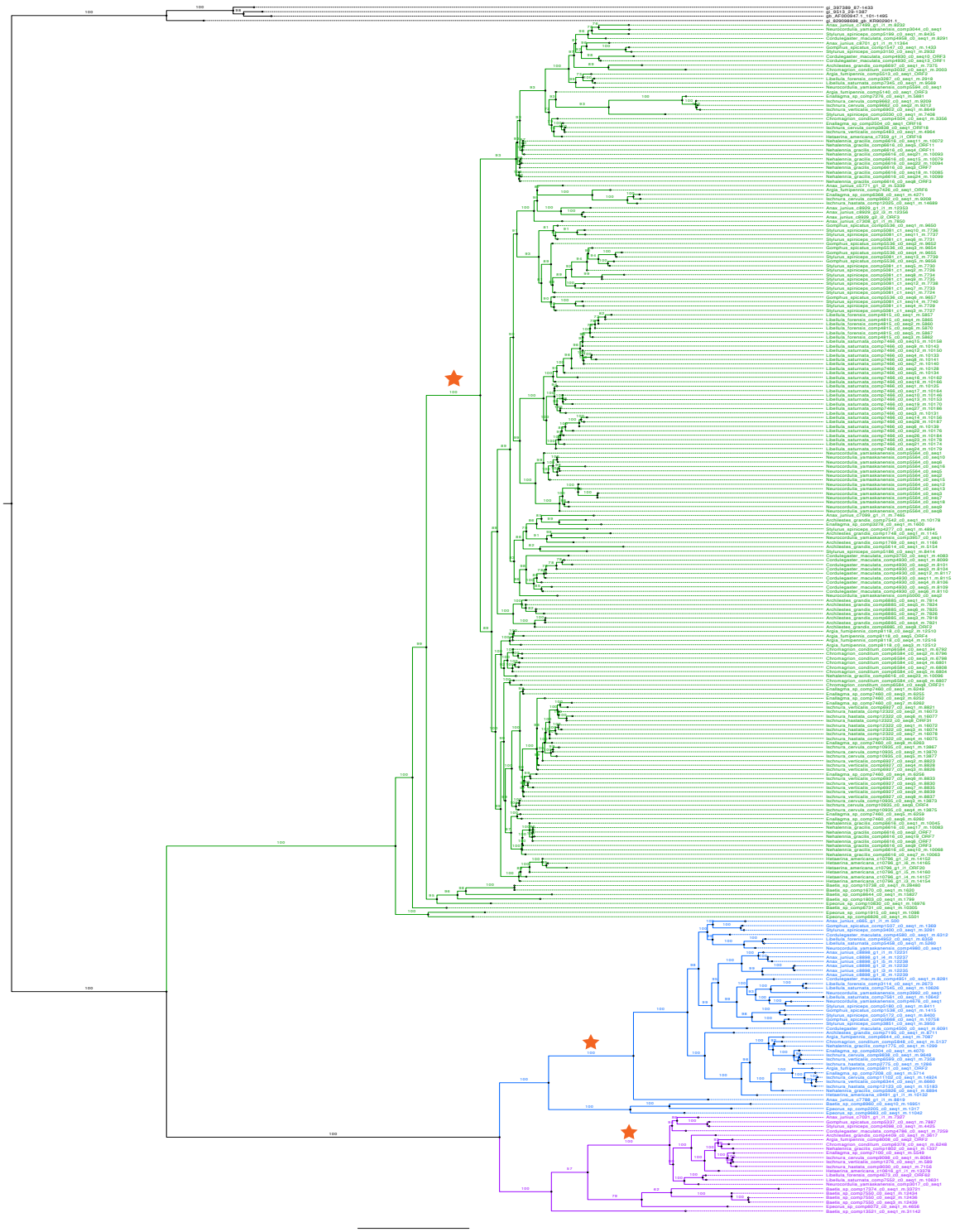


Figure S6. ML opsin tree. Opsin ancestral lineages are marked with the stars. Full length opsin sequences only (see Alignment 2).

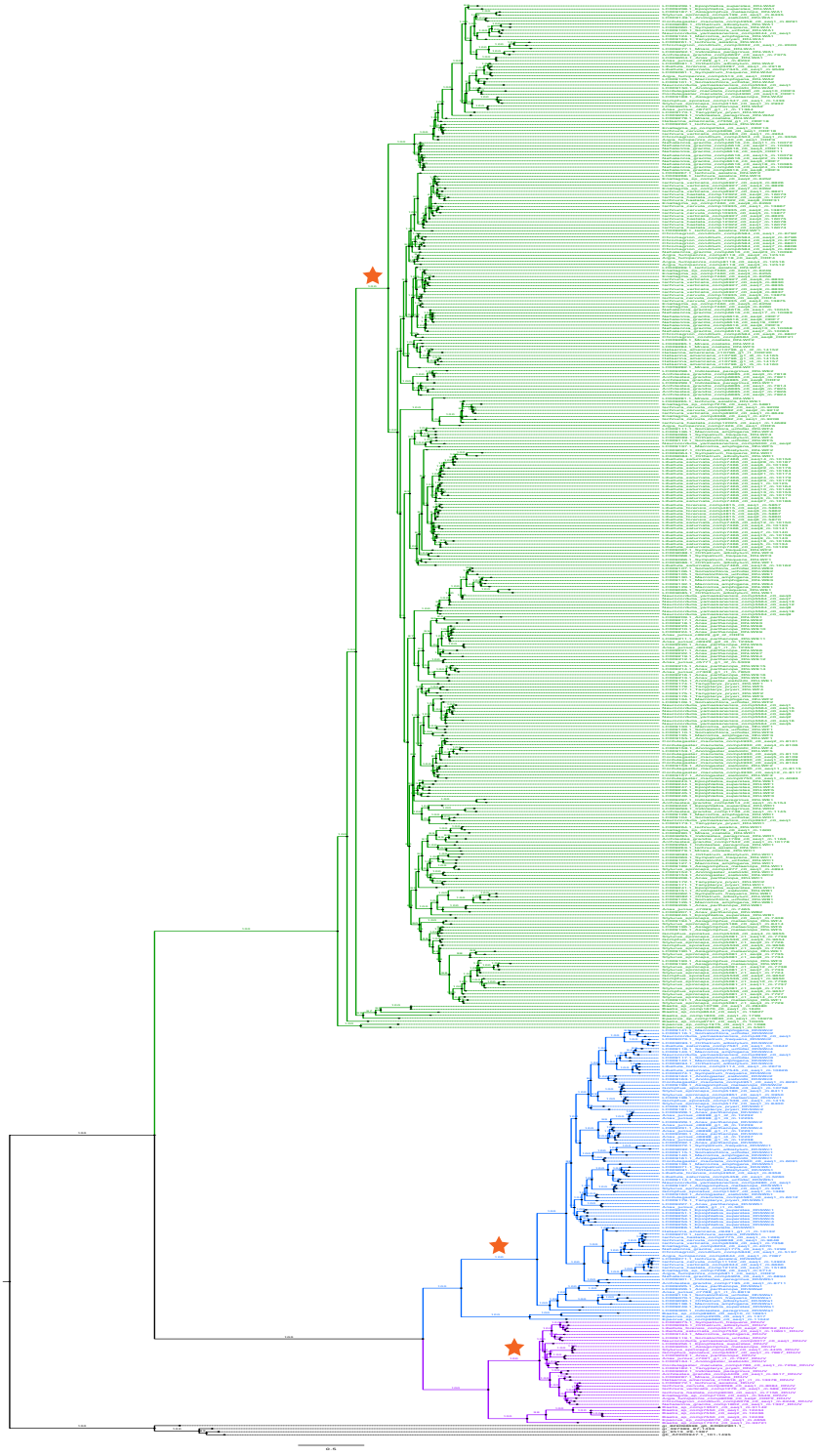


Figure S7. ML opsin tree. Opsin ancestral lineages are marked with the stars. All opsin sequences concatenated with all opsin sequences from (Futahashi *et al.* 2015) (see Alignment 3).

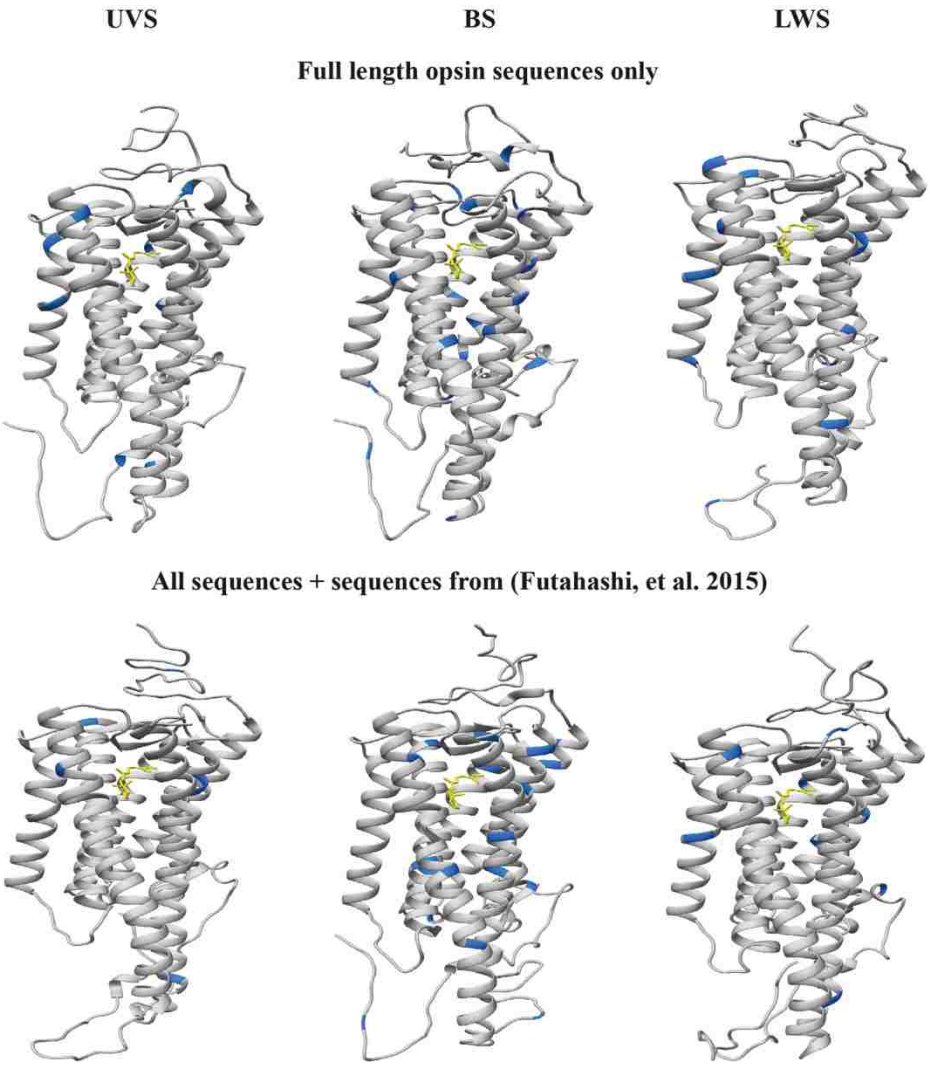


Figure S8. Predicted 3D models of reconstructed ancestral opsin proteins with positively selected sites (dN/dS (ω) > 1, blue) indentified by BEB algorithm (P > 0.95). Yellow structure represents a chromophore molecule. No sites within the binding pocket were found evolving under positive selection. Full-length opsin sequences only were used for inference (see Alignment 2) and all opsin sequences combined with all opsin sequences from (Futahashi *et al.* 2015) were used for inference (see Alignment 3).

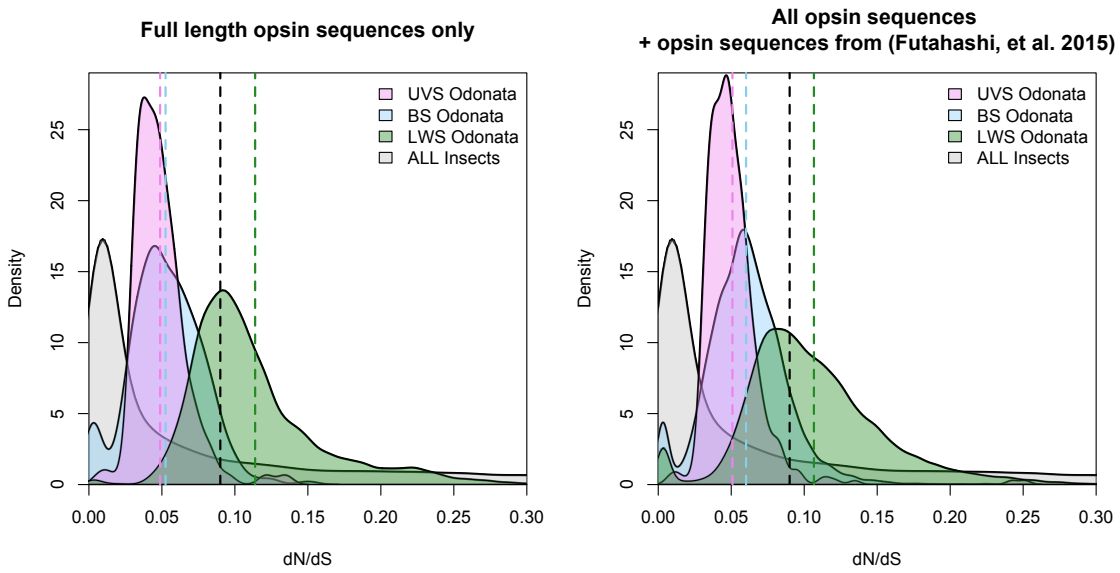


Figure S9. Distributions of pairwise dN/dS (ω) ratios calculated for each opsin class in Odonata and for all insect opsins. Dashed lines represent distribution means. Purifying selection is more relaxed in the LWS class than in the BS, UVS or all insect opsins.

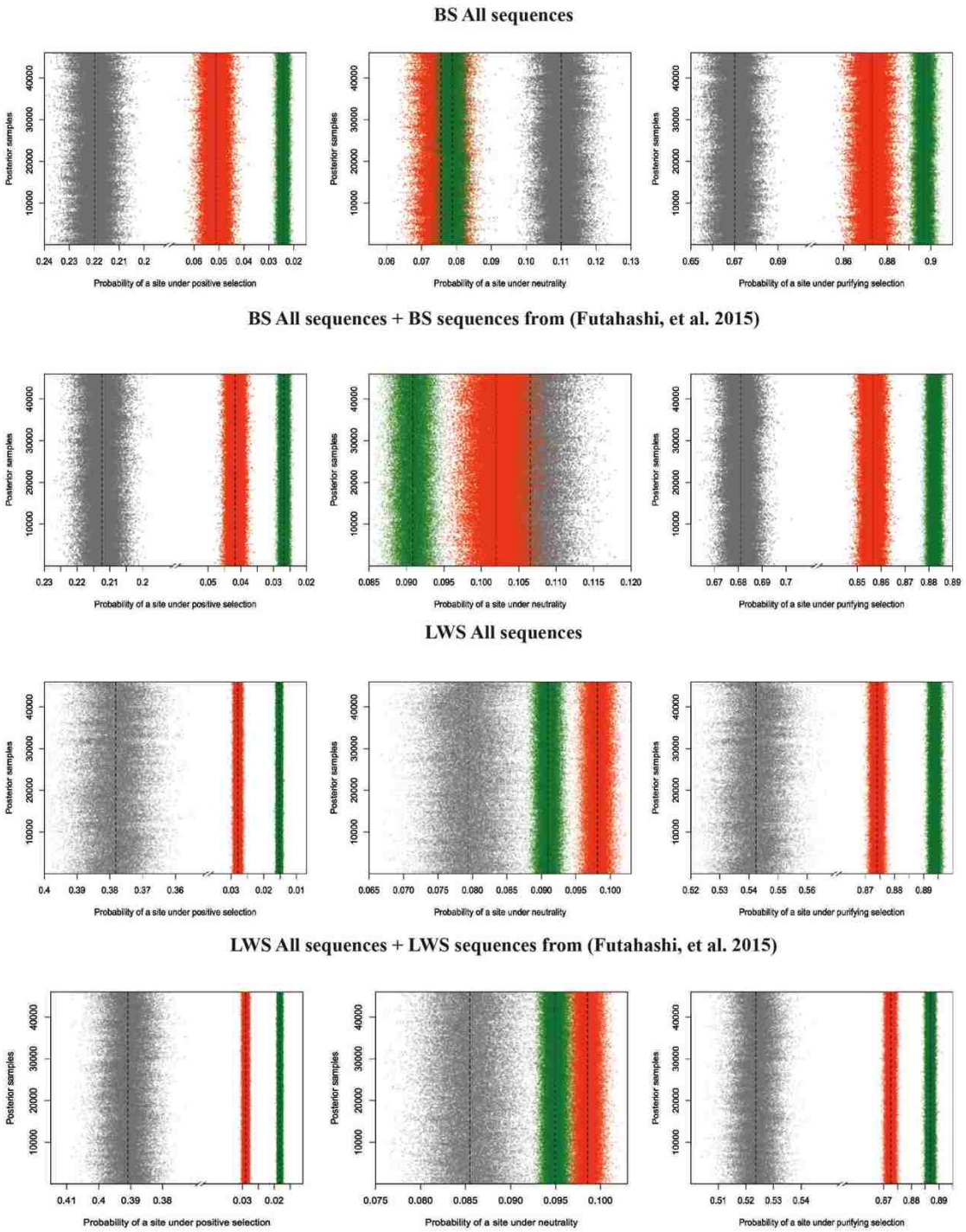


Figure S10. Posterior samples of probabilities of a site that is under positive, negative selection and neutrality before (grey), after duplication events (red) and terminal branches (green). The dashed lines represent posterior means.

Table S1. RNA-seq read library and assembly statistics.

Library	Reads Before Trimming	Reads After Trimming	N50	Max Contig Length	Min Contig Length	N of Contigs	N of Peptides (TransDecoder)	SRA ID (NCBI)
OD07_Cordulegaster_maculata	7207518	6819629	983	16508	201	28163	11877	SRR2164542
OD08_Anax_junius	6267456	5978119	807	9457	201	19987	8519	SRR2164543
OD10_Hetaerina_americana	6447244	6057989	1141	10815	201	34384	13373	SRR2164551
OD11_Ischnura_verticalis	6183018	5790475	879	7894	201	27001	10568	SRR2164552
OD12_Gomphus_spicatus	6498099	6273168	1502	11612	201	37936	10611	SRR2157378
OD13_Nehalennia_gracilis	5894197	5516510	1027	11774	201	33766	12694	SRR2157379
OD18_Chromagrion_conditum	7607629	7189745	1188	8671	201	30453	9256	SRR2157380
OD25_Stylurus_spiniceps	6840281	6597769	1249	23861	201	37436	13568	SRR2157381
OD28_Neurocordulia_yamaskanensis	6410925	6061801	1261	15978	201	34984	12905	SRR2157382
OD36_Argia_fumipennis_violacea	5955971	5600701	1076	11410	201	35049	12754	SRR2157383
OD42_Archilestes_grandis	8736454	8367974	1179	9315	201	34318	12285	SRR2164544
OD43_Hetaerina_americana_2	3585483	3411596	738	7343	201	24192	7318	SRR2164545
OD44_Enallagma_sp	5646110	5370720	940	6446	201	27135	8085	SRR2157367
OD45_Libellula_forensis	5591547	5383248	1125	13425	201	31962	11352	SRR2164546
OD46_Libellula_saturnata	5961628	5717528	1397	13986	201	35045	13326	SRR2164547
OD62_Ischnura_hastata	10080263	9551907	1777	11200	201	40154	13651	SRR2164548
OD64_Anax_junius_2	9657180	9195840	1133	21566	201	30833	13117	SRR2157371
OD_Ischnura_cervula	7105927	6702900	1156	15621	201	40741	14253	SRR2157372
R_E001_Baetis_sp	16352942	16113853	1786	10772	201	30517	16743	SRR2164549
R_E006_Epeorus_sp	13846765	13701079	1303	23453	201	45886	16782	SRR2164550

Table S2. Summary opsin table

Species	Library ID	Predicted Opsins	CD-HIT	Class	Length	Method	NCBI BLASTN Top Odonata Hit	NCBI ID	Futahashi Class	Stage	FPKM	Expression Test P Value FPKM	Pass Expression Test FPKM	Test Statisrtic (log(FPKM)) Distribution
Anax_junius	OD08_OD64	c5771_g1_i2|m.5339	KEEP	LWS	405	HMMER	Anax parthenope RhLWE12	LC009212.1	E	ADULT	19709.13	0.92634815	TRUE	SN(location=10.1068, scale=3.3935, slant=-7.2983)
Anax_junius	OD08_OD64	c665_g1_i1|m.500	KEEP	BS	555	HMMER	Anax parthenope RhSWb1	LC009227.1	b	ADULT	913.06	0.33229576	TRUE	
Anax_junius	OD08_OD64	c7021_g1_i1|m.7327	KEEP	UV	1164	HMMER	Anax parthenope RhUV	LC009233.1	NA	ADULT	2952.39	0.5328432	TRUE	
Anax_junius	OD08_OD64	c7093_g1_i1|m.7449	KEEP	LWS	891	BOTH	Anax parthenope RhLWE2	LC009217.1	E	ADULT	13.7	0.0273151	FALSE	
Anax_junius	OD08_OD64	c7093_g1_i2|m.7450	KEEP	LWS	813	BOTH	Anax parthenope RhLWE2	LC009217.1	E	ADULT	11.65	0.02414903	FALSE	
Anax_junius	OD08_OD64	c7093_g1_i3|m.7451	KEEP	LWS	687	HMMER	Anax parthenope RhLWE2	LC009217.1	E	ADULT	17.03	0.03212368	FALSE	
Anax_junius	OD08_OD64	c7093_g1_i4|m.7452	KEEP	LWS	891	BOTH	Anax parthenope RhLWE2	LC009217.1	E	ADULT	15.87	0.0304907	FALSE	
Anax_junius	OD08_OD64	c7099_g1_i1|m.7465	KEEP	LWS	885	BOTH	Anax parthenope RhLWB1	LC009206.1	B	LARVA	34.65	0.05316847	TRUE	
Anax_junius	OD08_OD64	c7308_g1_i1|m.7850	KEEP	LWS	438	HMMER	Anax parthenope RhLWE14	LC009214.1	E	ADULT	43735.44	0.99456741	TRUE	
Anax_junius	OD08_OD64	c7499_g1_i1|m.8232	KEEP	LWS	1134	BOTH	Anax parthenope RhLWA1	LC009204.1	A	LARVA	93.03	0.10048312	TRUE	
Anax_junius	OD08_OD64	c7788_g1_i1|m.8819	KEEP	BS	1134	BOTH	Anax parthenope RhSWa2	LC009226.1	a	LARVA	77.9	0.09010944	TRUE	
Anax_junius	OD08_OD64	c8701_g1_i1|m.11364	KEEP	LWS	996	BOTH	Anax parthenope RhLWA2	LC009205.1	A	ADULT	17606.23	0.90725966	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i1|m.12231	KEEP	BS	1143	HMMER	Anax parthenope RhSWc4	LC009231.1	c	ADULT	9582.22	0.78114084	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i2|m.12232	KEEP	BS	1056	BOTH	Anax parthenope RhSWc1	LC009228.1	c	ADULT	1657.99	0.42736706	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i3|m.12235	KEEP	BS	1065	BOTH	Anax parthenope RhSWc2	LC009229.1	c	ADULT	2341.35	0.48893454	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i4|m.12237	KEEP	BS	1143	BOTH	Anax parthenope RhSWc3	LC009230.1	c	ADULT	7920.44	0.7389781	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i5|m.12238	KEEP	BS	843	HMMER	Anax parthenope RhSWc3	LC009230.1	c	ADULT	5872.09	0.67368937	TRUE	
Anax_junius	OD08_OD64	c8898_g1_i6|m.12239	KEEP	BS	1137	BOTH	Anax parthenope RhSWc2	LC009229.1	c	ADULT	4607.77	0.6223617	TRUE	
Anax_junius	OD08_OD64	c8929_g1_i1|m.12353	KEEP	LWS	804	HMMER	Anax parthenope RhLWE5	LC009220.1	E	ADULT	24116.7	0.95473194	TRUE	
Anax_junius	OD08_OD64	c8929_g2_i1|m.12354	REMOVE	LWS	1131	BOTH	Anax parthenope RhLWE9	LC009224.1	E	ADULT	27851	0.97004202	TRUE	
Anax_junius	OD08_OD64	c8929_g2_i3|m.12356	KEEP	LWS	1128	BOTH	Anax parthenope RhLWE11	LC009211.1	E	ADULT	31813.41	0.98061179	TRUE	
Anax_junius	OD08_OD64	c8929_g2_i2|ORF3	KEEP	LWS	1209	PIA	Anax parthenope RhLWE9	LC009224.1	E	ADULT	27703.33	0.96955001	TRUE	
Archilestes_grandis	OD42	comp13860_c0_seq1|m.14353	KEEP	LWS	426	HMMER	Ischnura asiatica RhLWF1	LC009266.1	F	ADULT	7.37	0.01686301	FALSE	
Archilestes_grandis	OD42	comp14675_c0_seq1|m.14747	KEEP	LWS	657	BOTH	Macromia amphigena RhLWD1	LC009128.1	D	ADULT	6.34	0.01493127	FALSE	
Archilestes_grandis	OD42	comp15628_c0_seq1|m.15204	KEEP	LWS	369	HMMER	NA	NA	NA	ADULT	6.69	0.01559719	FALSE	
Archilestes_grandis	OD42	comp1748_c0_seq1|m.1145	KEEP	LWS	639	HMMER	Indolestes peregrinus RhLWD2	LC009296.1	D	ADULT	87.48	0.0967853	TRUE	
Archilestes_grandis	OD42	comp1769_c0_seq1|m.1166	KEEP	LWS	456	HMMER	Indolestes peregrinus RhLWD1	LC009295.1	D	ADULT	478.92	0.24619204	TRUE	
Archilestes_grandis	OD42	comp2634_c0_seq1|m.1945	KEEP	BS	480	HMMER	Indolestes peregrinus RhSWa1	LC009300.1	a	LARVA	5.5	0.01328914	FALSE	
Archilestes_grandis	OD42	comp4409_c0_seq1|m.3617	KEEP	UVS	1164	BOTH	Indolestes peregrinus RhUV	LC009302.1	NA	ADULT	1807.22	0.44230391	TRUE	
Archilestes_grandis	OD42	comp5614_c0_seq1|m.5154	KEEP	LWS	1278	BOTH	Indolestes peregrinus RhLWE1	LC009297.1	E	ADULT	287.31	0.19012076	TRUE	
Archilestes_grandis	OD42	comp6697_c0_seq1|m.7375	KEEP	LWS	726	HMMER	Indolestes peregrinus RhLWA1	LC009292.1	A	LARVA	434.35	0.23466263	TRUE	
Archilestes_grandis	OD42	comp6885_c0_seq1|m.7814	KEEP	LWS	1116	BOTH	Indolestes peregrinus RhLWF1	LC009299.1	F	ADULT	14116.55	0.86470859	TRUE	

Archilestes_grandis	OD42	comp6885_c0_seq2|m.7815	REMOVE	LWS	1122	BOTH	Indolestes peregrinus RhLWE2	LC009298.1	E	ADULT	17391.32	0.90505958	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq3|m.7818	KEEP	LWS	1122	BOTH	Indolestes peregrinus RhLWE2	LC009298.1	E	ADULT	24298.85	0.9556353	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq4|m.7821	KEEP	LWS	1122	BOTH	Indolestes peregrinus RhLWE2	LC009298.1	E	ADULT	22909.02	0.94826027	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq5|m.7824	KEEP	LWS	1131	BOTH	Indolestes peregrinus RhLWF1	LC009299.1	F	ADULT	8770.79	0.76153895	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq6|m.7825	KEEP	LWS	1131	BOTH	Indolestes peregrinus RhLWF1	LC009299.1	F	ADULT	12732.46	0.84316035	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq7|m.7826	KEEP	LWS	1116	BOTH	Indolestes peregrinus RhLWF1	LC009299.1	F	ADULT	18574.56	0.91657875	TRUE
Archilestes_grandis	OD42	comp7195_c0_seq1|m.8711	KEEP	BS	1149	BOTH	Indolestes peregrinus RhSWb1	LC009301.1	b	ADULT	430.7	0.23368485	TRUE
Archilestes_grandis	OD42	comp7542_c0_seq1|m.10178	KEEP	LWS	597	HMMER	Indolestes peregrinus RhLWD1	LC009295.1	D	ADULT	385.88	0.22121585	TRUE
Archilestes_grandis	OD42	comp6885_c0_seq8|ORF2	KEEP	LWS	1158	PIA	Indolestes peregrinus RhLWF1	LC009299.1	F	ADULT	19461.07	0.92431197	TRUE
Argia_fumipennis	OD36	comp2756_c0_seq1|m.2192	KEEP	LWS	639	HMMER	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	12.93	0.02614635	FALSE
Argia_fumipennis	OD36	comp5811_c0_seq1|m.5641	REMOVE	BS	1140	HMMER	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	155.55	0.13595188	TRUE
Argia_fumipennis	OD36	comp6644_c0_seq1|m.7087	KEEP	BS	1152	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	2860.33	0.52672933	TRUE
Argia_fumipennis	OD36	comp8008_c0_seq1|m.11583	REMOVE	UVS	1164	BOTH	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1327.66	0.39024131	TRUE
Argia_fumipennis	OD36	comp8118_c0_seq2|m.12510	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	20531.94	0.93272694	TRUE
Argia_fumipennis	OD36	comp8118_c0_seq3|m.12512	KEEP	LWS	1113	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	27506.15	0.96888021	TRUE
Argia_fumipennis	OD36	comp8118_c0_seq4|m.12516	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	21665.33	0.94063511	TRUE
Argia_fumipennis	OD36	comp5140_c0_seq1|ORF3	KEEP	LWS	888	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	LARVA	33239.64	0.98340285	TRUE
Argia_fumipennis	OD36	comp5513_c0_seq1|ORF2	KEEP	LWS	840	PIA	Sympetrum frequens RhLWA2	LC009061.1	A	LARVA	31.77	0.0500986	TRUE
Argia_fumipennis	OD36	comp5811_c0_seq1|ORF2	KEEP	BS	1254	PIA	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	155.55	0.13595188	TRUE
Argia_fumipennis	OD36	comp7426_c0_seq1|ORF6	KEEP	LWS	834	PIA	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	84.39	0.09467439	TRUE
Argia_fumipennis	OD36	comp8008_c0_seq2|ORF2	KEEP	UVS	1230	PIA	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1316.2	0.38883365	TRUE
Argia_fumipennis	OD36	comp8118_c0_seq5|ORF4	KEEP	LWS	756	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	14384.52	0.86854236	TRUE
Chromagrion_conditum	OD18	comp1840_c0_seq1|m.1242	KEEP	LWS	762	BOTH	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	26.08	0.04366656	FALSE
Chromagrion_conditum	OD18	comp3032_c0_seq1|m.2003	KEEP	LWS	303	HMMER	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	39.07	0.05767349	TRUE
Chromagrion_conditum	OD18	comp4504_c0_seq1|m.3356	KEEP	LWS	912	BOTH	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	22741.29	0.94729147	TRUE
Chromagrion_conditum	OD18	comp5848_c0_seq1|m.5137	KEEP	BS	1149	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	2932.77	0.53155337	TRUE
Chromagrion_conditum	OD18	comp6378_c0_seq1|m.6248	KEEP	UVS	1167	BOTH	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1073.74	0.35666849	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq1|m.6792	KEEP	LWS	1113	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	8804.69	0.76239361	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq2|m.6796	KEEP	LWS	1128	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	10333.7	0.79782453	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq3|m.6798	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	8042.84	0.74236501	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq4|m.6801	KEEP	LWS	1119	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	9380.16	0.77642133	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq5|m.6804	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	9951.69	0.78950983	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq6|m.6807	KEEP	LWS	1020	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	20663.49	0.93369538	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq7|m.6808	KEEP	LWS	1131	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	8962.51	0.76632997	TRUE
Chromagrion_conditum	OD18	comp6584_c0_seq8|ORF21	KEEP	LWS	831	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	13116.34	0.84944243	TRUE
Cordulegaster_maculata	OD07	comp3750_c0_seq1|m.4083	KEEP	LWS	468	HMMER	Anotogaster sieboldii	LC009157.1	F	ADULT	10934.02	0.81023773	TRUE

							RhLWF3						
Cordulegaster_maculata	OD07	comp4500_c0_seq1|m.6091	KEEP	BS	1161	BOTH	Anotogaster sieboldii RhSWc1	LC009161.1	c	ADULT	705.11	0.29571818	TRUE
Cordulegaster_maculata	OD07	comp4580_c0_seq1|m.6312	KEEP	BS	1128	HMMER	Anotogaster sieboldii RhSWb1	LC009160.1	b	ADULT	483.4	0.24731137	TRUE
Cordulegaster_maculata	OD07	comp4786_c0_seq1|m.7259	KEEP	UVS	1164	BOTH	Anotogaster sieboldii RhUV	LC009164.1	NA	ADULT	333.66	0.20545965	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq1|m.8099	KEEP	LWS	1128	BOTH	Anotogaster sieboldii RhLWF1	LC009155.1	F	ADULT	8224.82	0.74731081	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq11|m.8115	KEEP	LWS	1119	BOTH	Anotogaster sieboldii RhLWF2	LC009156.1	F	ADULT	5054.77	0.64176787	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq12|m.8117	KEEP	LWS	1122	BOTH	Anotogaster sieboldii RhLWF3	LC009157.1	F	ADULT	5353.61	0.65392982	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq2|m.8101	KEEP	LWS	1125	BOTH	Anotogaster sieboldii RhLWF1	LC009155.1	F	ADULT	7597.36	0.72979499	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq3|m.8104	KEEP	LWS	1131	BOTH	Anotogaster sieboldii RhLWF3	LC009157.1	F	ADULT	5227.93	0.64888874	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq4|m.8106	KEEP	LWS	1116	BOTH	Anotogaster sieboldii RhLWF1	LC009155.1	F	ADULT	7113.41	0.71533419	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq5|m.8109	KEEP	LWS	1128	BOTH	Anotogaster sieboldii RhLWF1	LC009155.1	F	ADULT	5532.02	0.66091147	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq6|m.8110	KEEP	LWS	1119	BOTH	Anotogaster sieboldii RhLWF4	LC009158.1	F	ADULT	4208.58	0.60361668	TRUE
Cordulegaster_maculata	OD07	comp4951_c0_seq1|m.8281	KEEP	BS	1149	BOTH	Anotogaster sieboldii RhSWc3	LC009163.1	c	ADULT	3797.32	0.58265751	TRUE
Cordulegaster_maculata	OD07	comp4958_c0_seq1|m.8291	KEEP	LWS	990	BOTH	Anotogaster sieboldii RhLWA1	LC009149.1	A	LARVA	3811.09	0.5833895	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq10|ORF3	KEEP	LWS	981	PIA	Anotogaster sieboldii RhLWF2	LC009156.1	F	ADULT	2162.47	0.47434667	TRUE
Cordulegaster_maculata	OD07	comp4930_c0_seq13|ORF1	KEEP	LWS	1026	PIA	Anotogaster sieboldii RhLWA2	LC009150.1	A	ADULT	3531.52	0.56807368	TRUE
Enallagma_sp	OD44	comp14216_c0_seq1|m.9644	KEEP	LWS	303	HMMER	Ischnura asiatica RhLWA1	LC009261.1	A	ADULT	9.68	0.02092484	FALSE
Enallagma_sp	OD44	comp3278_c0_seq1|m.1600	KEEP	LWS	399	HMMER	Ischnura asiatica RhLWD1	LC009264.1	D	ADULT	461.8	0.24185003	TRUE
Enallagma_sp	OD44	comp6204_c0_seq1|m.4070	KEEP	BS	1155	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	3187.53	0.54778004	TRUE
Enallagma_sp	OD44	comp6368_c0_seq1|m.4271	KEEP	LWS	438	BOTH	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	101.24	0.10575007	TRUE
Enallagma_sp	OD44	comp7100_c0_seq1|m.5549	KEEP	UVS	1167	BOTH	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	2050.66	0.46473694	TRUE
Enallagma_sp	OD44	comp7208_c0_seq1|m.5714	KEEP	BS	1137	BOTH	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	944.63	0.33731531	TRUE
Enallagma_sp	OD44	comp7276_c0_seq1|m.5881	KEEP	LWS	363	HMMER	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	171.79	0.14380338	TRUE
Enallagma_sp	OD44	comp7460_c0_seq1|m.6249	KEEP	LWS	1131	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	13524.26	0.8558558	TRUE
Enallagma_sp	OD44	comp7460_c0_seq2|m.6252	KEEP	LWS	1119	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	16997.6	0.90089849	TRUE
Enallagma_sp	OD44	comp7460_c0_seq3|m.6255	KEEP	LWS	516	HMMER	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	16811.24	0.89886856	TRUE
Enallagma_sp	OD44	comp7460_c0_seq4|m.6256	KEEP	LWS	1137	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	14907.67	0.87573228	TRUE
Enallagma_sp	OD44	comp7460_c0_seq5|m.6259	KEEP	LWS	801	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	26344.37	0.96461923	TRUE
Enallagma_sp	OD44	comp7460_c0_seq6|m.6260	KEEP	LWS	390	HMMER	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	15758.7	0.88663761	TRUE
Enallagma_sp	OD44	comp7460_c0_seq7|m.6262	KEEP	LWS	1125	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	20393.31	0.93169138	TRUE
Enallagma_sp	OD44	comp7460_c0_seq8|m.6263	KEEP	LWS	519	HMMER	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	25176.72	0.95974224	TRUE
Enallagma_sp	OD44	comp2504_c0_seq1|ORF16	KEEP	LWS	813	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	32068.5	0.98114436	TRUE
Gomphus_spicatus	OD12	comp1507_c0_seq1|m.1369	KEEP	BS	1143	BOTH	Asiagomphus melaenops RhSWb1	LC009197.1	b	ADULT	291.11	0.19143355	TRUE
Gomphus_spicatus	OD12	comp1538_c0_seq1|m.1415	KEEP	BS	1155	BOTH	Asiagomphus melaenops RhSWc1	LC009198.1	c	ADULT	5989.23	0.67793993	TRUE
Gomphus_spicatus	OD12	comp1547_c0_seq1|m.1433	KEEP	LWS	1134	BOTH	Asiagomphus melaenops RhLWA2	LC009188.1	A	LARVA	6021.73	0.67910618	TRUE

Gomphus_spicatus	OD12	comp5337_c0_seq1|m.7887	KEEP	UVS	1167	BOTH	Asiagomphus melaenops RhUV	LC009200.1	NA	ADULT	2791.17	0.52203008	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq1|m.9650	KEEP	LWS	1125	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	7742.74	0.73397178	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq2|m.9652	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	12226.81	0.83450399	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq3|m.9654	KEEP	LWS	1137	BOTH	Asiagomphus melaenops RhLWF6	LC009196.1	F	ADULT	9255.55	0.77345875	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq4|m.9655	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF5	LC009195.1	F	ADULT	10645.48	0.80436801	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq5|m.9656	KEEP	LWS	1137	BOTH	Asiagomphus melaenops RhLWF6	LC009196.1	F	ADULT	5649.55	0.6654038	TRUE
Gomphus_spicatus	OD12	comp5536_c0_seq6|m.9657	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	6208.94	0.68571725	TRUE
Gomphus_spicatus	OD12	comp5668_c0_seq1|m.10758	KEEP	BS	387	HMMER	Asiagomphus melaenops RhSWc2	LC009199.1	c	ADULT	645.18	0.28379434	TRUE
Hetaerina_americana	OD10_OD43	c10616_g1_i1|m.13378	KEEP	UVS	1164	BOTH	Mnais costalis RhUV	LC009287.1	NA	ADULT	1533.94	0.41414475	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i2|m.14152	KEEP	LWS	1008	BOTH	Mnais costalis RhLWF3	LC009284.1	F	ADULT	14361.47	0.86821668	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i3|m.14154	KEEP	LWS	1128	BOTH	Mnais costalis RhLWF2	LC009283.1	F	ADULT	16594.9	0.89646222	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i4|m.14157	KEEP	LWS	1131	BOTH	Mnais costalis RhLWF2	LC009283.1	F	ADULT	13144.72	0.84989725	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i5|m.14160	KEEP	LWS	1122	BOTH	Mnais costalis RhLWF2	LC009283.1	F	ADULT	12692.94	0.84249965	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i6|m.14165	KEEP	LWS	1125	BOTH	Mnais costalis RhLWF3	LC009284.1	F	ADULT	13027.76	0.84801443	TRUE
Hetaerina_americana	OD10_OD43	c13678_g1_i1|m.16021	KEEP	BS	309	HMMER	Anax parthenope RhSWc4	LC009231.1	c	ADULT	5.71	0.01370579	FALSE
Hetaerina_americana	OD10_OD43	c9491_g1_i1|m.10132	KEEP	BS	1155	BOTH	Mnais costalis RhSWb1	LC009286.1	b	ADULT	650.66	0.2849155	TRUE
Hetaerina_americana	OD10_OD43	c7359_g1_i1|ORF18	KEEP	LWS	888	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	23333.44	0.9506334	TRUE
Hetaerina_americana	OD10_OD43	c10796_g1_i1|ORF20	KEEP	LWS	804	PIA	Macromia amphigena RhLWF3	LC009135.1	F	ADULT	12176.84	0.83362421	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq1|m.13867	KEEP	LWS	1122	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	9419.93	0.77735842	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq2|m.13870	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	4611	0.62250763	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq3|m.13873	KEEP	LWS	1128	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	4955.7	0.63759823	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq4|m.13875	KEEP	LWS	1134	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	7258.38	0.7197591	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq5|m.13877	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	3996.48	0.59303437	TRUE
Ischnura_cervula	OD_I	comp11102_c0_seq1|m.14924	KEEP	BS	1137	BOTH	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	237.88	0.17198404	TRUE
Ischnura_cervula	OD_I	comp9098_c0_seq1|m.8084	KEEP	UVS	1167	BOTH	Ischnura asiatica RhUV	LC009272.1	UV	ADULT	835.8	0.31946716	TRUE
Ischnura_cervula	OD_I	comp9662_c0_seq1|m.9208	KEEP	LWS	609	HMMER	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	115.39	0.1143262	TRUE
Ischnura_cervula	OD_I	comp9662_c0_seq1|m.9209	KEEP	LWS	363	BOTH	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	115.39	0.1143262	TRUE
Ischnura_cervula	OD_I	comp9662_c0_seq2|m.9212	KEEP	LWS	363	BOTH	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	100.36	0.10519644	TRUE
Ischnura_cervula	OD_I	comp9838_c0_seq1|m.9648	KEEP	BS	1155	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	1058.41	0.35446192	TRUE
Ischnura_cervula	OD_I	comp3838_c0_seq1|ORF18	KEEP	LWS	912	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	10998.48	0.81152528	TRUE
Ischnura_cervula	OD_I	comp10935_c0_seq6|ORF4	KEEP	LWS	1179	PIA	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	3980.57	0.59222159	TRUE
Ischnura_hastata	OD62	comp12025_c0_seq1|m.14689	KEEP	LWS	921	BOTH	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	188.68	0.15153287	TRUE
Ischnura_hastata	OD62	comp12123_c0_seq1|m.15183	KEEP	BS	1305	BOTH	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	95.18	0.10188498	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq1|m.16072	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	8548.8	0.75586079	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq2|m.16073	KEEP	LWS	1128	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	6827.18	0.70634954	TRUE

Ischnura_hastata	OD62	comp12322_c0_seq3|m.16074	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	11909.41	0.82883915	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq4|m.16075	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	6940.92	0.70996023	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq5|m.16076	REMOVE	LWS	1122	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	5260.89	0.6502209	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq6|m.16077	KEEP	LWS	1122	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	6710.15	0.7025769	TRUE
Ischnura_hastata	OD62	comp12322_c0_seq7|m.16078	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	6795.31	0.70532801	TRUE
Ischnura_hastata	OD62	comp15900_c0_seq1|m.20216	KEEP	BS	315	HMMER	Anax parthenope RhSWc5	LC009232.1	c	ADULT	8.32	0.01857499	FALSE
Ischnura_hastata	OD62	comp17538_c0_seq1|m.21045	KEEP	LWS	399	HMMER	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	6.95	0.01608542	FALSE
Ischnura_hastata	OD62	comp2775_c0_seq1|m.1286	KEEP	BS	1155	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	472.36	0.24454052	TRUE
Ischnura_hastata	OD62	comp9030_c0_seq1|m.7156	KEEP	UVS (fuse)	369	HMMER	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1274.81	0.38367225	TRUE
Ischnura_hastata	OD62	comp9312_c0_seq1|m.7691	KEEP	UVS (fuse)	819	HMMER	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	856.24	0.32294036	TRUE
Ischnura_hastata	OD62	comp9591_c0_seq1|m.8139	KEEP	LWS	501	HMMER	Anax parthenope RhLWE5	LC009220.1	E	ADULT	13.15	0.02648268	FALSE
Ischnura_hastata	OD62	comp12322_c0_seq8|ORF31	KEEP	LWS	1149	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	4689.61	0.62603237	TRUE
Ischnura_verticalis	OD11	comp1276_c0_seq1|m.589	KEEP	UVS (fuse)	813	HMMER	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	5835.45	0.67234444	TRUE
Ischnura_verticalis	OD11	comp5483_c0_seq1|m.4964	KEEP	LWS	867	HMMER	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	54323.8	0.99813628	TRUE
Ischnura_verticalis	OD11	comp5892_c0_seq1|m.5644	KEEP	UVS(fuse)	366	BOTH	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1888.35	0.45002821	TRUE
Ischnura_verticalis	OD11	comp6344_c0_seq1|m.6660	KEEP	BS	1137	BOTH	Ischnura asiatica RhSWb2	LC009271.1	b	ADULT	663.01	0.28741869	TRUE
Ischnura_verticalis	OD11	comp6599_c0_seq1|m.7358	KEEP	BS	1155	BOTH	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	4382.32	0.61195351	TRUE
Ischnura_verticalis	OD11	comp6902_c0_seq1|m.8649	KEEP	LWS	363	HMMER	Ischnura asiatica RhLWE1	LC009265.1	E	ADULT	97.29	0.1032449	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq1|m.8821	KEEP	LWS	1125	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	23454	0.95128754	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq2|m.8823	KEEP	LWS	1116	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	20932.28	0.93563187	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq3|m.8826	KEEP	LWS	1119	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	18494.38	0.91584275	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq4|m.8828	KEEP	LWS	1122	BOTH	Ischnura asiatica RhLWF3	LC009268.1	F	ADULT	25801.31	0.96242976	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq5|m.8830	KEEP	LWS	1128	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	16661.96	0.89721391	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq6|m.8833	KEEP	LWS	1137	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	14078.79	0.86415994	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq7|m.8835	KEEP	LWS	693	HMMER	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	10492.59	0.80118582	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq8|m.8837	KEEP	LWS	1134	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	19716.93	0.92641131	TRUE
Ischnura_verticalis	OD11	comp6927_c0_seq9|m.8839	KEEP	LWS	1131	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	12466.41	0.83866107	TRUE
Libellula_forensis	OD45	comp3114_c0_seq1|m.2673	KEEP	BS	1143	BOTH	Orthetrum albistylum RhSWc3	LC009094.1	c	ADULT	751.09	0.30439965	TRUE
Libellula_forensis	OD45	comp3287_c0_seq1|m.2918	KEEP	LWS	918	BOTH	Orthetrum albistylum RhLWA2	LC009081.1	A	ADULT	10510.62	0.80156373	TRUE
Libellula_forensis	OD45	comp3358_c0_seq1|m.3025	KEEP	BS	1146	BOTH	Orthetrum albistylum RhSWc2	LC009093.1	c	ADULT	20.84	0.03721725	FALSE
Libellula_forensis	OD45	comp4673_c0_seq1|m.5456	REMOVE	UVS	1173	BOTH	Orthetrum albistylum RhUV	LC009095.1	NA	ADULT	710.29	0.2967155	TRUE
Libellula_forensis	OD45	comp4815_c0_seq1|m.5857	KEEP	LWS	1206	BOTH	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	5564.12	0.66214667	TRUE
Libellula_forensis	OD45	comp4815_c0_seq2|m.5860	KEEP	LWS	1158	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	5671.08	0.66621784	TRUE
Libellula_forensis	OD45	comp4815_c0_seq3|m.5862	KEEP	LWS	1212	BOTH	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	5204.68	0.64794463	TRUE
Libellula_forensis	OD45	comp4815_c0_seq4|m.5865	KEEP	LWS	1158	BOTH	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	5360.68	0.65421033	TRUE
Libellula_forensis	OD45	comp4815_c0_seq5|m.5867	KEEP	LWS	1164	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	5333.2	0.65311822	TRUE

Libellula_forensis	OD45	comp4815_c0_seq6|m.5870	KEEP	LWS	1206	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	5545.45	0.66142902	TRUE
Libellula_forensis	OD45	comp4952_c0_seq1|m.6358	KEEP	BS	1137	BOTH	Orthetrum albistylum RhSWb1	LC009091.1	b	ADULT	95.4	0.1020275	TRUE
Libellula_forensis	OD45	comp4673_c0_seq2|ORF62	KEEP	UVS	1173	PIA	Orthetrum albistylum RhUV	LC009095.1	NA	ADULT	708.74	0.2964176	TRUE
Libellula_saturnata	OD46	comp13280_c0_seq1|m.14510	KEEP	BS	381	HMMER	Orthetrum albistylum RhSWb1	LC009091.1	b	ADULT	12.58	0.02560718	FALSE
Libellula_saturnata	OD46	comp5458_c0_seq1|m.5260	KEEP	BS	804	HMMER	Orthetrum albistylum RhSWb1	LC009091.1	b	ADULT	34.73	0.05325215	TRUE
Libellula_saturnata	OD46	comp5990_c0_seq1|m.6128	KEEP	LWS	1143	BOTH	Orthetrum albistylum RhLWA1	LC009080.1	A	LARVA	18.1	0.03359398	FALSE
Libellula_saturnata	OD46	comp7345_c0_seq1|m.9569	KEEP	LWS	1146	BOTH	Orthetrum albistylum RhLWA2	LC009081.1	A	ADULT	12492.02	0.83909946	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq1|m.10125	KEEP	LWS	1230	HMMER	Orthetrum albistylum RhLWF4	LC009089.1	F	ADULT	385.31	0.22105143	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq10|m.10146	KEEP	LWS	1113	HMMER	Orthetrum albistylum RhLWF4	LC009089.1	F	ADULT	483.71	0.24738857	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq12|m.10150	KEEP	LWS	1224	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	1103.12	0.36083404	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq13|m.10153	KEEP	LWS	1221	BOTH	Orthetrum albistylum RhLWF4	LC009089.1	F	ADULT	464.13	0.24244711	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq14|m.10156	KEEP	LWS	1110	BOTH	Orthetrum albistylum RhLWD1	LC009084.1	D	ADULT	505.5	0.25273511	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq15|m.10158	KEEP	LWS	1230	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	730.62	0.30058167	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq16|m.10162	KEEP	LWS	972	HMMER	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	481.33	0.24679503	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq17|m.10164	KEEP	LWS	1122	BOTH	Orthetrum albistylum RhLWF4	LC009089.1	F	ADULT	432.91	0.23427752	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq18|m.10166	KEEP	LWS	1230	BOTH	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	611.52	0.27676109	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq19|m.10170	KEEP	LWS	1128	HMMER	Orthetrum albistylum RhLWF4	LC009089.1	F	ADULT	439.72	0.23609152	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq2|m.10128	KEEP	LWS	1113	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	845.8	0.3211738	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq21|m.10174	KEEP	LWS	1113	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	464.64	0.24257754	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq22|m.10176	KEEP	LWS	990	HMMER	Orthetrum albistylum RhLWF3	LC009088.1	F	ADULT	364.36	0.21489719	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq23|m.10178	KEEP	LWS	1116	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	821.35	0.3169754	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq24|m.10179	KEEP	LWS	1122	BOTH	Orthetrum albistylum RhLWF3	LC009088.1	F	ADULT	383.05	0.22039796	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq26|m.10184	KEEP	LWS	1083	HMMER	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	273.88	0.18539385	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq27|m.10186	KEEP	LWS	1116	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	892.5	0.3289597	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq28|m.10187	KEEP	LWS	1119	BOTH	Orthetrum albistylum RhLWD1	LC009084.1	D	ADULT	430.64	0.23366874	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq3|m.10131	KEEP	LWS	1224	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	845.84	0.32118059	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq4|m.10133	KEEP	LWS	1131	BOTH	Orthetrum albistylum RhLWF3	LC009088.1	F	ADULT	961.53	0.33995311	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq5|m.10134	KEEP	LWS	1221	HMMER	Orthetrum albistylum RhLWF1	LC009086.1	F	ADULT	783.04	0.31021629	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq6|m.10139	KEEP	LWS	1113	BOTH	Orthetrum albistylum RhLWD1	LC009084.1	D	ADULT	850.69	0.32200315	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq7|m.10140	KEEP	LWS	1116	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	1030.19	0.35033848	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq8|m.10141	KEEP	LWS	1137	BOTH	Orthetrum albistylum RhLWF3	LC009088.1	F	ADULT	562.04	0.26592852	TRUE
Libellula_saturnata	OD46	comp7466_c0_seq9|m.10143	KEEP	LWS	1122	BOTH	Orthetrum albistylum RhLWF2	LC009087.1	F	ADULT	757.27	0.30553805	TRUE
Libellula_saturnata	OD46	comp7545_c0_seq1|m.10626	KEEP	BS	1290	HMMER	Orthetrum albistylum RhSWc3	LC009094.1	c	ADULT	1597.11	0.42097589	TRUE
Libellula_saturnata	OD46	comp7552_c0_seq1|m.10631	KEEP	UVS	1173	HMMER	Orthetrum albistylum RhUV	LC009095.1	NA	ADULT	611.07	0.2766653	TRUE
Libellula_saturnata	OD46	comp7561_c0_seq1|m.10642	KEEP	BS	1146	HMMER	Orthetrum albistylum RhSWc2	LC009093.1	c	ADULT	640.37	0.28280487	TRUE

Nehalennia_gracilis	OD13	comp10758_c0_seq1|m.13974	KEEP	LWS	381	HMMER	Asiagomphus melaenops RhLWF1	LC009191.1	F	ADULT	22.77	0.03965997	FALSE
Nehalennia_gracilis	OD13	comp1775_c0_seq1|m.1299	KEEP	BS	1152	HMMER	Ischnura asiatica RhSWb1	LC009270.1	b	ADULT	5077.12	0.64269875	TRUE
Nehalennia_gracilis	OD13	comp1802_c0_seq1|m.1337	KEEP	UVS	1167	HMMER	Ischnura asiatica RhUV	LC009272.1	NA	ADULT	1942	0.45499564	TRUE
Nehalennia_gracilis	OD13	comp25998_c0_seq1|m.20879	KEEP	BS	417	HMMER	Asiagomphus melaenops RhSWc1	LC009198.1	c	ADULT	8.28	0.01850414	FALSE
Nehalennia_gracilis	OD13	comp5926_c0_seq1|m.6894	KEEP	BS	1152	HMMER	Mnais costalis RhSWb1	LC009286.1	b	ADULT	582.72	0.27053134	TRUE
Nehalennia_gracilis	OD13	comp6038_c0_seq2|m.7146	KEEP	LWS	1266	BOTH	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	10.85	0.02286314	FALSE
Nehalennia_gracilis	OD13	comp6038_c0_seq3|m.7149	KEEP	LWS	951	HMMER	Ischnura asiatica RhLWA1	LC009261.1	A	LARVA	5.98	0.01423537	FALSE
Nehalennia_gracilis	OD13	comp6616_c0_seq1|m.10045	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	3195.73	0.54828439	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq10|m.10068	KEEP	LWS	1131	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	6279.01	0.68814608	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq11|m.10072	KEEP	LWS	1110	BOTH	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2018.37	0.46188521	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq13|m.10075	REMOVE	LWS	1104	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	3944.24	0.59035534	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq15|m.10079	KEEP	LWS	1119	BOTH	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2455.64	0.49779704	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq17|m.10083	KEEP	LWS	1110	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	4556.47	0.62003224	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq18|m.10085	KEEP	LWS	1131	BOTH	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2196.38	0.47718364	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq21|m.10093	KEEP	LWS	1104	HMMER	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2538.67	0.50403081	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq22|m.10094	KEEP	LWS	1113	BOTH	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2973.92	0.53425058	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq23|m.10096	KEEP	LWS	978	BOTH	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	31134.11	0.97911684	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq24|m.10099	KEEP	LWS	1125	BOTH	Mnais costalis RhLWA2	LC009278.1	A	ADULT	2903.01	0.52958342	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq7|m.10063	KEEP	LWS	1125	BOTH	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	4687.09	0.62592017	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq2|ORF7	KEEP	LWS	1137	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	2451.03	0.49744601	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq3|ORF7	KEEP	LWS	1197	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	2307.06	0.48620837	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq4|ORF11	KEEP	LWS	1137	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	2397.89	0.49336102	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq5|ORF11	KEEP	LWS	1137	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	2304	0.48596353	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq6|ORF7	KEEP	LWS	1137	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	4224.21	0.60437852	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq8|ORF3	KEEP	LWS	1218	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	2371.3	0.49128979	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq9|ORF3	KEEP	LWS	1218	PIA	Ischnura asiatica RhLWF4	LC009269.1	F	ADULT	4302.39	0.60815366	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq19|ORF7	KEEP	LWS	1131	PIA	Ischnura asiatica RhLWF2	LC009267.1	F	ADULT	3398.12	0.56040662	TRUE
Nehalennia_gracilis	OD13	comp6616_c0_seq26|ORF7	REMOVE	LWS	1137	PIA	Ischnura asiatica RhLWA2	LC009262.1	A	ADULT	1722.69	0.43396511	TRUE
Neurocordulia_yamaskanensis	OD28	comp3017_c0_seq1|m.2863	KEEP	UVS	1170	HMMER	Somatochlora uchidai RhUV	LC009119.1	NA	ADULT	746.2	0.30349424	TRUE
Neurocordulia_yamaskanensis	OD28	comp3044_c0_seq1|m.2896	KEEP	LWS	639	HMMER	Somatochlora uchidai RhLWA1	LC009100.1	A	LARVA	133.17	0.12433921	TRUE
Neurocordulia_yamaskanensis	OD28	comp3957_c0_seq1|m.4147	KEEP	LWS	771	HMMER	Somatochlora uchidai RhLWD1	LC009104.1	D	ADULT	318.15	0.20048327	TRUE
Neurocordulia_yamaskanensis	OD28	comp3992_c0_seq1|m.4205	KEEP	BS	1308	HMMER	Somatochlora uchidai RhSWc4	LC009118.1	c	ADULT	6403.33	0.69239635	TRUE
Neurocordulia_yamaskanensis	OD28	comp4676_c0_seq1|m.5412	KEEP	BS	468	HMMER	Somatochlora uchidai RhSWc2	LC009116.1	c	ADULT	105.5	0.10839517	TRUE
Neurocordulia_yamaskanensis	OD28	comp4980_c0_seq1|m.6265	KEEP	BS	1137	HMMER	Somatochlora uchidai RhSWb1	LC009114.1	b	ADULT	290.19	0.19111669	TRUE
Neurocordulia_yamaskanensis	OD28	comp5000_c0_seq2|m.6322	KEEP	LWS	630	HMMER	Somatochlora uchidai RhLWF5	LC009112.1	F	ADULT	6254.35	0.68729406	TRUE

Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq1|m.9119	KEEP	LWS	1332	BOTH	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	5166.79	0.64639817	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq10|m.9143	KEEP	LWS	1338	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	3104.22	0.54259453	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq11|m.9147	REMOVE	LWS	1122	BOTH	Somatochlora uchidai RhLWE3	LC009107.1	E	ADULT	3222.85	0.54994488	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq12|m.9148	KEEP	LWS	1329	BOTH	Somatochlora uchidai RhLWE3	LC009107.1	E	ADULT	1173.83	0.37053662	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq13|m.9151	KEEP	LWS	1110	BOTH	Somatochlora uchidai RhLWE3	LC009107.1	E	ADULT	1778.24	0.43947964	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq15|m.9156	KEEP	LWS	1113	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	5746.77	0.66905821	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq16|m.9159	KEEP	LWS	942	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	4577.79	0.62100306	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq18|m.9163	KEEP	LWS	1338	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	2732.78	0.51798847	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq2|m.9122	KEEP	LWS	1335	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	4575.94	0.62091897	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq3|m.9125	KEEP	LWS	1122	BOTH	Somatochlora uchidai RhLWE3	LC009107.1	E	ADULT	1836.25	0.44509815	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq5|m.9129	KEEP	LWS	1116	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	5816.41	0.6716426	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq6|m.9131	KEEP	LWS	1119	HMMER	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	2643.15	0.5116464	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq8|m.9135	KEEP	LWS	1332	BOTH	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	2248.55	0.48148138	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq9|m.9139	KEEP	LWS	1335	BOTH	Somatochlora uchidai RhLWF2	LC009109.1	F	ADULT	4234.51	0.60487926	TRUE
Neurocordulia_yamaskanensis	OD28	comp5594_c0_seq1|m.9421	KEEP	LWS	405	HMMER	Somatochlora uchidai RhLWA2	LC009101.1	A	ADULT	9911.86	0.78862341	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq4|ORF3	REMOVE	LWS	1326	PIA	Somatochlora uchidai RhLWE2	LC009106.1	F	ADULT	1884.87	0.44970222	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq7|ORF7	KEEP	LWS	1146	PIA	Somatochlora uchidai RhLWE2	LC009106.1	F	ADULT	2443.18	0.49684705	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq14|ORF24	REMOVE	LWS	93	PIA	Somatochlora uchidai RhLWE2	LC009106.1	F	ADULT	1660.4	0.42761634	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq17|ORF24	REMOVE	LWS	108	PIA	Somatochlora uchidai RhLWE2	LC009106.1	F	ADULT	932.69	0.33543121	TRUE
Neurocordulia_yamaskanensis	OD28	comp5564_c0_seq19|ORF24	REMOVE	LWS	108	PIA	Somatochlora uchidai RhLWE2	LC009106.1	F	ADULT	1232.26	0.37823467	TRUE
Stylurus_spiniceps	OD25	comp3150_c0_seq1|m.2932	KEEP	LWS	1005	BOTH	Asiagomphus melaenops RhLWA2	LC009188.1	A	ADULT	5320.38	0.65260706	TRUE
Stylurus_spiniceps	OD25	comp3400_c0_seq1|m.3281	KEEP	BS	1143	HMMER	Asiagomphus melaenops RhSWb1	LC009197.1	b	ADULT	398.43	0.22479733	TRUE
Stylurus_spiniceps	OD25	comp3851_c0_seq1|m.3950	KEEP	BS	396	HMMER	Asiagomphus melaenops RhSWc2	LC009199.1	c	ADULT	1948.33	0.45557465	TRUE
Stylurus_spiniceps	OD25	comp4098_c0_seq1|m.4425	KEEP	UVS	1167	HMMER	Asiagomphus melaenops RhUV	LC009200.1	NA	ADULT	1028.34	0.35006531	TRUE
Stylurus_spiniceps	OD25	comp4277_c0_seq1|m.4894	KEEP	LWS	1140	BOTH	Asiagomphus melaenops RhLWC1	LC009189.1	C	LARVA	84.58	0.09480531	TRUE
Stylurus_spiniceps	OD25	comp5030_c0_seq1|m.7408	KEEP	LWS	321	HMMER	Asiagomphus melaenops RhLWE1	LC009190.1	E	ADULT	2424.52	0.49541707	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq1|m.7724	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	5524.39	0.66061695	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq10|m.7736	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF5	LC009195.1	F	ADULT	3505.1	0.56657454	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq11|m.7737	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF1	LC009191.1	F	ADULT	4749.9	0.62870149	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq12|m.7738	KEEP	LWS	1134	HMMER	Asiagomphus melaenops RhLWF2	LC009192.1	F	ADULT	4122.47	0.59937601	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq13|m.7739	KEEP	LWS	1134	BOTH	Asiagomphus melaenops RhLWF5	LC009195.1	F	ADULT	9877.19	0.78784877	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq14|m.7740	KEEP	LWS	1134	BOTH	Asiagomphus melaenops RhLWF1	LC009191.1	F	ADULT	2903.89	0.52964191	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq2|m.7726	KEEP	LWS	1137	BOTH	Asiagomphus melaenops RhLWF6	LC009196.1	F	ADULT	4616.49	0.62275546	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq3|m.7727	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	4537.27	0.61915462	TRUE
Stylurus_spiniceps	OD25	comp5081_c1_seq4|m.7729	KEEP	LWS	1134	BOTH	Asiagomphus melaenops RhLWF1	LC009191.1	F	ADULT	4103.82	0.59844768	TRUE

Stylurus_spiniceps	OD25	comp5081_c1_seq5|m.7730	KEEP	LWS	1137	BOTH	Asiagomphus melaenops RhLWF6	LC009196.1	F	ADULT	7762.96	0.73454687	TRUE	
Stylurus_spiniceps	OD25	comp5081_c1_seq6|m.7731	KEEP	LWS	1128	BOTH	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	5913.21	0.67518993	TRUE	
Stylurus_spiniceps	OD25	comp5081_c1_seq7|m.7733	KEEP	LWS	1134	HMMER	Asiagomphus melaenops RhLWF3	LC009193.1	F	ADULT	3332.43	0.55653905	TRUE	
Stylurus_spiniceps	OD25	comp5081_c1_seq8|m.7734	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWE1	LC009190.1	E	ADULT	4046	0.59554679	TRUE	
Stylurus_spiniceps	OD25	comp5081_c1_seq9|m.7735	KEEP	LWS	1131	BOTH	Asiagomphus melaenops RhLWE1	LC009190.1	E	ADULT	3533.77	0.56820091	TRUE	
Stylurus_spiniceps	OD25	comp5172_c0_seq1|m.8400	KEEP	BS	1155	HMMER	Asiagomphus melaenops RhSWc1	LC009198.1	c	ADULT	5239.22	0.64934587	TRUE	
Stylurus_spiniceps	OD25	comp5180_c0_seq1|m.8411	KEEP	BS	744	HMMER	Asiagomphus melaenops RhSWc2	LC009199.1	c	ADULT	2565.72	0.50602634	TRUE	
Stylurus_spiniceps	OD25	comp5186_c0_seq1|m.8414	KEEP	LWS	471	HMMER	Asiagomphus melaenops RhLWF4	LC009194.1	F	ADULT	1506.08	0.4110639	TRUE	
Stylurus_spiniceps	OD25	comp5199_c0_seq1|m.8435	KEEP	LWS	1008	BOTH	Asiagomphus melaenops RhLWA1	LC009187.1	A	LARVA	828.93	0.31828632	TRUE	
Baetis_sp	R_EP_001	comp10738_c0_seq1|m.28480	KEEP	LWS	594	HMMER	NA	NA	NA	NA	286.88	0.85627769	TRUE	SN(location=3.5203 , scale=2.0106, slant=0)
Baetis_sp	R_EP_001	comp13521_c0_seq1|m.31142	KEEP	UVS	420	HMMER	NA	NA	NA	NA	2.58	0.10036508	TRUE	
Baetis_sp	R_EP_001	comp1670_c0_seq1|m.1620	KEEP	LWS	549	HMMER	NA	NA	NA	NA	425.41	0.89611125	TRUE	
Baetis_sp	R_EP_001	comp17374_c0_seq1|m.33721	KEEP	UVS	372	HMMER	NA	NA	NA	NA	2.04	0.08131587	TRUE	
Baetis_sp	R_EP_001	comp1803_c0_seq1|m.1799	KEEP	LWS	840	HMMER	NA	NA	NA	NA	70.84	0.64360474	TRUE	
Baetis_sp	R_EP_001	comp23897_c0_seq1|m.36105	KEEP	LWS	348	HMMER	NA	NA	NA	NA	0.95	0.037835	FALSE	
Baetis_sp	R_EP_001	comp6731_c0_seq1|m.10305	KEEP	LWS	1140	BOTH	NA	NA	NA	NA	706.18	0.93470415	TRUE	
Baetis_sp	R_EP_001	comp7550_c0_seq1|m.12434	KEEP	UVS	1140	BOTH	NA	NA	NA	NA	12.85	0.31528342	TRUE	
Baetis_sp	R_EP_001	comp7550_c0_seq2|m.12436	KEEP	UVS	1157	BOTH	NA	NA	NA	NA	8.77	0.25113393	TRUE	
Baetis_sp	R_EP_001	comp7550_c0_seq3|m.12439	KEEP	UVS	933	BOTH	NA	NA	NA	NA	8.58	0.24767658	TRUE	
Baetis_sp	R_EP_001	comp8644_c0_seq1|m.15827	KEEP	LWS	1098	BOTH	NA	NA	NA	NA	80.92	0.66795701	TRUE	
Baetis_sp	R_EP_001	comp8960_c0_seq10|m.16951	KEEP	BS	1152	BOTH	NA	NA	NA	NA	8.11	0.23890202	TRUE	
Epeorus_sp	R_EP_006	comp10830_c0_seq1|m.16976	KEEP	LWS	1131	BOTH	NA	NA	NA	NA	754.94	0.93882454	TRUE	
Epeorus_sp	R_EP_006	comp1915_c0_seq1|m.1098	KEEP	LWS	1116	BOTH	NA	NA	NA	NA	6.35	0.20284122	TRUE	
Epeorus_sp	R_EP_006	comp2205_c0_seq1|m.1317	KEEP	BS	654	HMMER	NA	NA	NA	NA	4.41	0.1555672	TRUE	
Epeorus_sp	R_EP_006	comp6072_c0_seq1|m.4656	KEEP	UVS	1176	BOTH	NA	NA	NA	NA	11.17	0.29094849	TRUE	
Epeorus_sp	R_EP_006	comp6826_c0_seq1|m.5501	KEEP	LWS	1131	BOTH	NA	NA	NA	NA	646.37	0.92891487	TRUE	
Epeorus_sp	R_EP_006	comp9683_c0_seq1|m.11042	KEEP	BS	1143	BOTH	NA	NA	NA	NA	33.37	0.49749146	TRUE	

Table S3. Analyses of Positive Selection using Branch-Site Models (PAML) on UVS, BS and LWS Ancestral Branches of ML Opsin Gene Tree. Full length opsins only (see Alignment 2). Positively selected site numbers may differ between Alignment 1 and Alignment 3 due to alignment variability.

Branch[a]	Model	LnL	Site Class	Proportion (p)	Background (ω)	Foreground (ω)	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	A	-67126.516977	0	0.78274	0.07070	0.07070	22, 86, 127, 197, 204, 234, 287, 326, 327, 399	
			1	0.13481	1	1		
			2a	0.07034	0.07070	17.61057		
			2b	0.01211	1	17.61057		
	Null	-67132.712716	0	0.75702	0.07056	0.07056		
			1	0.13019	1	1		
			2a	0.09624	0.07056	1		
			2b	0.01655	1	1		
								0.00043
BS	A	-67119.630357	0	0.63932	0.07174	0.07174	61, 64, 90, 95, 101, 118, 134, 142, 183, 211, 227, 243, 253, 276, 305, 310, 323, 336, 342, 357, 372, 377, 415	
			1	0.10975	1	1		
			2a	0.21416	0.07174	1		
			2b	0.03676	72.19361	72.19361		
	Null	-67129.745031	0	0.60537	0.07133	0.07133		
			1	0.10424	1	1		
			2a	0.24774	0.07133	1		
			2b	0.04266	1	1		
								$<10^{-5}$
UVS	A	-67127.604164	0	0.78544	0.07262	0.07262	25, 43, 98, 141, 184, 197, 231, 239, 284, 291, 303, 341, 342, 372, 413	
			1	0.12798	1	1		
			2a	0.07445	0.07262	36.95341		
			2b	0.01213	1	36.95341		
	Null	-67132.480289	0	0.73315	0.07233	0.07233		
			1	0.11940	1	1		
			2a	0.12680	0.07233	1		
			2b	0.02065	1	1		
								0.00179

Table S4. Analyses of Positive Selection using Branch-Site Models (PAML) on UVS, BS and LWS Ancestral Branches of ML Opsin Gene Tree. All opsin sequences concatenated with all opsin sequences from (Futahashi *et al.* 2015) (see Alignment 3). Positively selected site numbers may differ between Alignment 1 and Alignment 2 due to alignment variability.

Branch[a]	Model	LnL	Site Class	Proportion (p)	Background (ω)	Foreground (ω)	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	A	-151694.45908	0	0.76806	0.06996	0.06996	91, 97, 138, 208, 245, 300, 302, 338, 339, 392	
			1	0.13553	1	1		
			2a	0.08195	0.06996	39.72358		
			2b	0.01446	1	39.72358		
	Null	-151702.11696	0	0.73266	0.06987	0.06987		
			1	0.12811	1	1		
			2a	0.11851	0.06987	1		
			2b	0.02072	1	1		
								0.00009
BS	A	-151692.57478	0	0.61394	0.07044	0.07044	26, 72, 75, 101, 112, 129, 145, 164, 238, 254, 264, 287, 317, 322, 335, 348, 354, 369, 389, 425	
			1	0.11016	1	1		
			2a	0.23393	0.07044	999		
			2b	0.04197	1	999		
	Null	-151700.33599	0	0.62166	0.07024	0.07024		
			1	0.10991	1	1		
			2a	0.22810	0.07024	1		
			2b	0.04033	1	1		
								0.00008
UVS	A	-151705.20325	0	0.77006	0.07024	0.07024	26, 54, 152, 250, 298, 353, 354, 384	
			1	0.13346	1	1		
			2a	0.08222	0.07024	23.48868		
			2b	0.01425	1	23.48868		
	Null	-151708.5164	0	0.71073	0.07014	0.07014		
			1	0.12671	1	1		
			2a	0.13797	0.07014	1		
			2b	0.02460	1	1		
								0.01

Table S5. Analyses of Positive Selection using Random Site Models (PAML) on UVS, BS and LWS ML opsin trees. Full length opsins only (see Alignment 2).

Opsin Class	Site Model	LnL	Site Class	Proportion (p)	ω	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	M0	-34035.26865	0	NA	0.10076	NA	
	M1a	-32511.23902	0	0.83217	0.04867	NA	
			1	0.16783	1		
	M2a	-32476.61274	0	0.83043	0.04824	3, 21, 45, 375, 379	$< 10^{-10}$ (M2a vs. M1a)
			1	0.16957	1		
			2	0.00001	7.70889		
	M3	-31894.35685	0	0.61202	0.01176		$< 10^{-10}$ (M3 vs. M0)
			1	0.28249	0.17326		
			2	0.10549	0.86092		
	M7	-31822.53884	0-9	0-9:0.1	0 0.00004 0.00067 0.00440 0.01796 0.05486 0.13728 0.29415 0.54829 0.87418	NA	
	M8	-31750.62259	0-10	0-9: 0.09657, 10: 0.03429	0 0.00017 0.00148 0.00620 0.01815 0.04311 0.08975 0.17192 0.31593 0.59740 1.56938	3, 21, 45, 375, 379	$< 10^{-10}$ (M8 vs. M7)
BS	M0	-16536.68486	0	NA	0.08311	NA	
	M1a	-16320.04099	0	0.88879	0.06290	NA	
			1	0.11121	1		
	M2a	-16320.04099	0	0.88879	0.06290		> 0.99 (M2a vs. M1a)
			1	0.00019	1		
			2	0.11102	1		
	M3	-15985.3611	0	0.41704	0.00149	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.38985	0.07928		
			2	0.19311	0.34095		
	M7	-15982.37403	0-9	0-9:0.1	0.00003 0.00080 0.00397 0.01150 0.02570 0.04961 0.08779 0.14821 0.24881 0.45637	NA	
	M8	-15982.37425	0-10	0-9:0.1,10: 0.00001	0.00003 0.00080 0.00398 0.01150 0.02569 0.04961 0.08778 0.14820 0.24878 0.45633 1.000000.14732 0.25203 0.46868 1	NA	> 0.99 (M8 vs. M7)

UVS	M0	-6991.80722	0	NA	0.06167	NA	
	M1a	-6913.55876	0	0.93414	0.03948	NA	
			1	0.06586	1		
	M2a	-6913.55876	0	0.93415	0.03948		> 0.99 (M2a vs. M1a)
			1	0.00039	1		
			2	0.06546	1		
	M3	-6836.19231	0	0.70175	0.00269	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.26312	0.17383		
			2	0.03513	0.59783		
	M7	-6838.88333	0-9	0-9:0.1	0 0.00001 0.00011 0.00076 0.00335 0.01098 0.02986 0.07219 0.16517 0.40289	NA	
	M8	-6838.25805	0-10	0-9: 0.09941, 10: 0.00591	0 0.00001 0.00013 0.00086 0.00353 0.01098 0.02860 0.06688 0.14947 0.36258 1	NA	0.263 (M8 vs. M7)

Table S6. Analyses of Positive Selection using Random Site Models (PAML) on UVS, BS and LWS ML opsin trees. All opsin sequences concatenated with all opsin sequences from (Futahashi *et al.* 2015) (see Alignment 3).

Opsin Class	Site Model	LnL	Site Class	Proportion (p)	ω	Positively Selected Sites (BEB, P > 0.95)	LRT *P* value
LWS	M0	-34035.26865	0	NA	0.10076	NA	
	M1a	-32511.23902	0	0.83217	0.04867	NA	
			1	0.16783	1		
	M2a	-32476.61274	0	0.83043	0.04824	3, 21, 45, 375, 379	$< 10^{-10}$ (M2a vs. M1a)
			1	0.16957	1		
			2	0.00001	7.70889		
	M3	-31894.35685	0	0.61202	0.01176		$< 10^{-10}$ (M3 vs. M0)
			1	0.28249	0.17326		
			2	0.10549	0.86092		
	M7	-31822.53884	0-9	0-9:0.1	0 0.00004 0.00067 0.00440 0.01796 0.05486 0.13728 0.29415 0.54829 0.87418	NA	
	M8	-31750.62259	0-10	0-9: 0.09657, 10: 0.03429	0 0.00017 0.00148 0.00620 0.01815 0.04311 0.08975 0.17192 0.31593 0.59740 1.56938	3, 21, 45, 375, 379	$< 10^{-10}$ (M8 vs. M7)
BS	M0	-16536.68486	0	NA	0.08311	NA	
	M1a	-16320.04099	0	0.88879	0.06290	NA	
			1	0.11121	1		
	M2a	-16320.04099	0	0.88879	0.06290		> 0.99 (M2a vs. M1a)
			1	0.00019	1		
			2	0.11102	1		
	M3	-15985.3611	0	0.41704	0.00149	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.38985	0.07928		
			2	0.19311	0.34095		
	M7	-15982.37403	0-9	0-9:0.1	0.00003 0.00080 0.00397 0.01150 0.02570 0.04961 0.08779 0.14821 0.24881 0.45637	NA	
	M8	-15982.37425	0-10	0-9:0.1,10: 0.00001	0.00003 0.00080 0.00398 0.01150 0.02569 0.04961 0.08778 0.14820 0.24878	NA	0.984 (M8 vs. M7)

					0.45633 1.000000.14732 0.25203 0.46868 1		
UVS	M0	-6991.80722	0	NA	0.06167	NA	
	M1a	-6913.55876	0	0.93414	0.03948	NA	
			1	0.06586	1		
	M2a	-6913.55876	0	0.93415	0.03948		> 0.99 (M2a vs. M1a)
			1	0.00039	1		
			2	0.06546	1		
	M3	-6836.19231	0	0.70175	0.00269	NA	$< 10^{-10}$ (M3 vs. M0)
			1	0.26312	0.17383		
			2	0.03513	0.59783		
	M7	-6838.88333	0-9	0-9:0.1	0 0.00001 0.00011 0.00076 0.00335 0.01098 0.02986 0.07219 0.16517 0.40289	NA	
	M8	-6838.25805	0-10	0-9: 0.09941, 10: 0.00591	0 0.00001 0.00013 0.00086 0.00353 0.01098 0.02860 0.06688 0.14947 0.36258 1	NA	0.263 (M8 vs. M7)

Supplementary References

Futahashi R, Kawahara-Miki R, Kinoshita M*, et al.* (2015) Extraordinary diversity of visual opsin genes in dragonflies. *Proc Natl Acad Sci U S A* **112**, E1247-1256.

Soria-Carrasco V, Talavera G, Igea J, Castresana J (2007) The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics* **23**, 2954-2956.

Chapter 2

Transcriptomic Data Resolve the Phylogenetic Backbone for Odonata

Anton Suvorov*[§1], M. Stanley Fujimoto*[2], Paul Bodily[2], Mark Clement[2], Keith A. Crandall[3], Michael F. Whiting[1,4] and Seth M. Bybee[1,4]

[1]Department of Biology, Brigham Young University, Provo, UT
[2]Computer Science Department, Brigham Young University, Provo, UT
[3]Computational Biology Institute, George Washington University, Ashburn, VA
[4]M.L. Bean Museum, Brigham Young University, Provo, UT

*Equal contributors

§Corresponding author

Abstract

In order to reconstruct the evolutionary history of Odonata (dragonflies and damselflies), we performed comprehensive phylotranscriptomic analyses of 83 species covering 75% of all extant odonate families. Using maximum likelihood, Bayesian, coalescent-based and alignment free tree inference frameworks we were able to test, refine and resolve previously controversial relationships within the order. In particular, we confirmed the monophyly of Zygoptera, recovered Gomphidae and Petaluridae as sister groups with high confidence and identified Calopterygoidea as monophyletic. Our fossil calibration coupled with diversification analyses provided insight into key events that influenced the evolution of Odonata. Specifically, we determined that there was possible mass extinction of ancient odonte diversity during the P-Tr crisis, and a single odonate lineage persisted following this extinction event; the first major radiation events occurred during the Cretaceous followed by almost a ~70 Ma period of protracted speciation; second speciation burst happened after K-Pg crisis (which is consistent

with bird speciation patterns). We also evaluated various effects of accelerated evolution, missing data, taxon rougness, orthology detection and DNA vs. amino acid data types, partition schemes on topological stability and consistency of tree estimation. We found that higher substitution rates and the amount of data and type have the greatest impacts on stability and consistency respectively.

Introduction

Odonata, the insect order that contains dragonflies and damselflies, lacks a strongly supported backbone to clearly resolve higher-level phylogenetic relationships (Carle *et al.* 2015; Dijkstra *et al.* 2014a). Current data recover odonates together with Ephemeroptera (mayflies) as the *living* representatives of the most ancient insect lineages to have evolved wings and active flight (Thomas *et al.* 2013). Odonates possess unique anatomical and morphological features such as a specialized body form, specialized wing venation, a distinctive form of muscle attachment to the wing base (Busse *et al.* 2013) allowing for direct flight and accessory (secondary) male genitalia that support certain unique behaviors (e.g., sperm competition). They are among the most adept flyers of all animals and are exclusively carnivorous insects relying primarily on vision (Chauhan *et al.* 2014; Suvorov *et al.* 2017) to capture prey. They spend much of their adult life in flight (Corbet 1999). Biogeographically, odonates exhibit worldwide dispersal (Troast *et al.* 2016) and play crucial ecological roles in local freshwater communities, being a top invertebrate predator (Dijkstra *et al.* 2014b). Due to this combination of characteristics, odonates are quickly becoming model organisms to study specific questions in ecology, physiology and evolution (Bybee *et al.* 2016; Cordoba-Aguilar 2008).

Odonata is comprised of 39 families (11 in Anisoptera, 1 in Anisozygoptera and 27 in Zygoptera) and a handful of groups that are *incertae sedis* and approximately 6000 described species (Dijkstra *et al.* 2013). The modern order Odonata emerged ~268 Ma from the Upper Permian (Nel *et al.* 2011) and is divided into three monophyletic suborders: Zygoptera, Anisoptera and Anisozygoptera. The latter two suborders are often combined into a single suborder Epiprocta based on adult morphology (e.g., wing venation, positioning of the compound eyes, overall body plan, etc.) (Rehn 2003). Despite these derived characteristics

(synapomorphies), some phylogenetic estimates recovered non-monophyletic Zygoptera (estimates based on a single mitochondrial DNA 12S marker (Saux *et al.* 2003), nuclear 28S and mitochondrial 16S markers (Hasegawa & Kasuya 2006) or specific wing-based morphological (Trueman 1996)), whereas others identified Anisoptera and Zygoptera as monophyletic lineages estimates based on molecular data alone and morphological and molecular data combined (Bybee *et al.* 2008; Carle *et al.* 2008; Dumont *et al.* 2010; Suvorov *et al.* 2017) (fig. 1).

To date, the relationships among most superfamilies as well as families remain unresolved, receiving low phylogenetic support from different scale molecular and/or morphology-based studies (Odonata: (Bybee *et al.* 2008; Carle *et al.* 2008; Dumont *et al.* 2010; Hasegawa & Kasuya 2006; Saux *et al.* 2003) Anisoptera: (Carle *et al.* 2015; Misof *et al.* 2001; Ware *et al.* 2007) Zygoptera: (Dijkstra *et al.* 2014a; Dumont *et al.* 2005)). In particular, the relative position of Calopterygoidea (Dijkstra *et al.* 2014a), a zygopteran complex that includes most of the families of the suborder, is still controversial. Likewise, the sister group to Cavilabiata, an anisopteran group consisting of all families except Aeshnidae, Petaluridae and Gomphidae, remains unidentified.

The higher-level classification within Zygoptera is largely uncertain and remains undetermined for almost all families. In a recent comprehensive Sanger-based study using three genes (16S, 28S and COI) that included almost all extant zygopteran families Dijkstra *et al.* (2014a) reconstructed a pectinate phylogeny with weakly supported clades across the backbone of the topology, although family level groupings were well-supported. They identified the monophyly of Lestoidea, Platystictidae as sister to Calopterygoidea + Coenagrionoidea, and the monophyly of Platycnemididae + Coenagrionidae nested within paraphyletic Calopterygoidea (fig. 1B). Similar hypotheses were also proposed in (Bybee *et al.* 2008; Carle *et al.* 2008).

A higher-level phylogenetic classification of anisopteran families is more stable and well-supported compared to Zygoptera (Dijkstra *et al.* 2014a), nevertheless the relationships between some key families remain unresolved. In particular, it is unclear whether Gomphidae is sister to Aeshnidae (Carle *et al.* 2008), forms a monophyletic group with Petaluridae (Carle *et al.* 2015) or if Petaluridae is sister to Gomphidae + Cavilabiata (Bybee *et al.* 2008) (fig. 1C). The placement of Corduliidae has been disputed, as Ware *et al.* (2007) resolve Corduliidae as sister to Macromiidae + Libellulidae, but Carle *et al.* (2015) proposed a competing hypothesis of Corduliidae being sister to Libellulidae. In both cases, recovered phylogenetic relationships were weakly supported. Phylogenetic relationships derived from morphological data contradict the aforementioned hypotheses by reconstructing Gomphidae as sister to all remaining Anisoptera, as well as recovering a highly unresolved Corduliidae + Macromiidae + Synthemistidae + Libellulidae (Blanke *et al.* 2013).

Next-generation sequencing technologies (e.g., RNA-seq) provide a fast and cost-effective means to amass large amounts of "omic" data necessary for resolving phylogenetic relationships with high confidence (Breinholt *et al.* 2017; Misof *et al.* 2014). Here we generated and compiled RNA-seq data from 83 odonate species from 28 family level (~75% of families) including all three suborders (Bybee *et al.* 2016). Our taxon sampling coupled with an extensive gene data coverage provides us with a new perspective for odonate phylogeny reconstructed from the largest dataset to date. Using a large array of data types (DNA and amino acid), orthology-detection approaches and phylogenetic tools, we were able to robustly resolve an odonate backbone that will serve as a useful blueprint to further address fine-scale evolutionary relationships within Odonata. Thus, we were able to test multiple phylogenetic hypotheses established by previous authors (fig. 1)

Results and Discussion

Data Summary, Supermatrix Statistics and Phylogenetic Analyses

The accession numbers (will be provided upon acceptance/request), N50s and other information about each RNA-seq library are summarized in supplementary table 1. In the present study three types of homologous loci (gene clusters) namely conservative single-copy orthologs (CO), all single-copy orthologs (AO) and paralogy-parsed orthologs (PO) identified by BUSCO v1.22 (Simao *et al.* 2015), OrthoMCL (Li *et al.* 2003) and Yang's (Yang & Smith 2014) pipelines respectively were used in phylogenetic inference. Eventually, 1603 CO, 1643 AO and 4341 PO gene clusters with ≥ 42 (~50%) species present were used to develop our supermatrices. Each gene cluster was aligned, trimmed and concatenated (for more details, see Materials and Methods) resulting in five main supermatrices, CO, AO and PO, which included 2,167,861 DNA (682,327 amino acid (AA) sites), 882,417 AA sites, 6,202,646 DNA (1,605,370 AA) sites, respectively (for more details, see supplementary table 2). Four contrasting tree building methods (ML:IQTREE, Bayesian:ExaBayes, Supertree:ASTRAL, Alignment-Free (AF): Co-phylog) were used to infer odonate phylogenetic relationships using different input data types (untrimmed and trimmed supermatrices, codon supermatrices, codon supermatrices with 1st and 2nd or 3rd positions removed, gene trees and assembled transcriptomes). In total we performed 48 phylogenetic analyses and compared topologies to identify stable and conflicting relationships (supplementary table 2).

High-Level Phylogenetic Backbone of Odonata: Subordinal Relationships

Divergence of Zygoptera and Epiprocta (Anisozygoptera+Anisoptera) from the Most

Recent Common Ancestor (MRCA) occurred in the Middle Triassic ~242 Ma (fig. 2A) which is in line with recent estimates (Misof *et al.* 2014; Thomas *et al.* 2013). Comprehensive phylogenetic coestimation of subordinal relationships within Odonata showed that the suborders were well supported (fig. 3) being consistently recovered as monophyletic clades in all analyses. Monophyly of Anisoptera was rejected by only two analyses: the Bayesian analysis of the CO AA trimmed supermatrix due to the placement of *Epiophlebia superstes* within Anisoptera (which also failed to converge ASDSF = 46.983%, PSRF = 1.248, ESS = 20.677, see Materials and Methods), and the AF method, which is known to be sensitive to variation in the amount of phylogenetic data (Yi & Jin 2013) (further discussion of non-converged Bayesian and AF trees will be omitted as a result unless otherwise noted). The monopyhy of Zygoptera was supported in all analyses with the exception of the AF method that spuriously nested species across the entire phylogenetic tree. Generally, the topological disagreements between AF and other phylogenetic trees in terms of Robinson-Foulds (RF) distances attain its maximum (RF = 0.63) for the AF tree being equidistant from all other trees (fig. 4A). In a few previous studies, however, paraphyletic relationships of Zygoptera had been proposed based on wing vein characters derived from fossil odonatoids and extant Odonata (Trueman 1996), analysis of 12S (Saux *et al.* 2003), analysis of 18S, 28S, Histone 3 (H3) and morphological data (Ogden & Whiting 2003) and analysis of 16S and 28S dataset (Hasegawa & Kasuya 2006). In most of these analyses, Lestidae was sister to Anisoptera. Functional morphology comparisons of flight systems, secondary male genitalia and ovipositor also supported a non-monophyletic Zygoptera with the uncertain phylogenetic placement of multiple groups (Pfau 1991). Nevertheless, the relationships inferred from these previous datasets seem to be highly unlikely due to an apparent morphological differentiation (e.g., eye spacing, body robustness, wing shape) between the

suborders and support for monophyletic Anisoptera and Zygoptera from more recent morphological (Busse *et al.* 2015), molecular (Carle *et al.* 2008; Kim *et al.* 2014; Suvorov *et al.* 2017; Thomas *et al.* 2013) and combined studies using both data types (Bybee *et al.* 2008).

Diversification analysis

Investigation of diversification rates in Odonata highlighted two major trends correlated with two mass extinction events in the Permian-Triassic (P-Tr) ~252 Ma and Cretaceous-Paleogene (K-Pg) ~66 Ma. First, it appears likely that the ancient diversity of Odonata was largely eliminated after the P-Tr mass extinction event as was also the case for multiple insect lineages (Labandeira & Sepkoski 1993). According to the fossil record at least two major odonatoid lineages went extinct (Protodonata and Protanisoptera (Grimaldi & Engel 2005)) and likely many genera from other lineages as well (e.g., Kargalotypidae from Triadophlebimorpha (Nel *et al.* 2001)). The establishment of major odonate lineages was observed during the Cretaceous starting ~135 Ma (fig. 2A) which coincided with the radiation of angiosperm plants that in turn triggered formation of herbivorous insect lineages (Misof *et al.* 2014). Odonates are exclusively carnivorous insects and their diversification was likely driven by the aforementioned sequence of events. Interestingly, molecular adaptations in the odonate visual systems are coupled with their diversification during the Cretaceous as well (Suvorov *et al.* 2017). Second, we observed rapid cladogenesis of multiple odonate lineages following the K-Pg mass extinction (fig. 2A,B) which can be attributed to the "Pull of the Present", where lineages that evolved in the recent past are less likely to go extinct than lineages that arose in the distant past (Nee *et al.* 1994; Pybus & Harvey 2000). Indeed, the statistical analysis of the lineage through time (LTT) plot (fig. 2B) illuminates irregular patterns of species accumulation, rejecting constant birth-

death process by Constant-Rates (CR) test (Pybus & Harvey 2000) ($P = 0.0156$). A very similar pattern of species radiation was documented for birds (Prum *et al.* 2015). Interestingly, we identified γ-statistic < 0 (fig. 2B) which is indicative of lowered rates of death process resulting in nodes being closer to the tree root than expected accompanied by longer branch lengths (Pybus & Harvey 2000). This may suggest that odonates actually experienced protracted speciation, i.e. a slowdown in the rate of a branching process (Etienne & Rosindell 2012) during the Cretaceous. However some young odonate lineages, like Libellulidae or Coenagrionidae, exhibited more resent speciation events (internal nodes are closer to tree tips) during the Cenozoic, which indicates a complex pattern of irregular rate turnovers of cladogenesis throughout geological time across Odonata.

Substitution rate, missing data and clade stability

In order to evaluate clade stability (i.e., recovery of a clade with acceptable nodal support), we calculated three characteristics, namely substitution rates, amount of missing data and taxon "jumpiness" (a.k.a. wildcard/rogue taxon (Aberer *et al.* 2013)) (fig. 5). Correlation analyses showed that only missing data significantly negatively contributed to branch stability (Spearman's rank correlation, $P = 0.005$) which was also suggested by a previous simulation study (Lemmon *et al.* 2009). However, the impact of missing data has only minor influence on the accuracy of phylogenetic inference if a sufficient amount of informative characters are sampled (Wiens & Morrill 2011). We also note that while the removal of a potentially wildcard taxon may improve nodal support locally on the topology, it does not necessarily mean the position of the wildcard taxon on a tree was recovered entirely incorrectly. Additionally, we observed that branches under accelerated evolution (increased substitution rates and potential

saturation) pose challenges to phylogenetic estimation by affecting inferential methods so that they generate incongruent trees with conflicting topologies (fig. 3). In turn, slowly evolving branches were ubiquitously recovered across various methods. In particular, we identified three hotspots of accelerated evolution (figs. 3, 5), specifically, the branches of Calopterygoidea (1), Libellulidae (2) and Petaluridae+Gomphidae (3) (fig. 5), on the odonate phylogeny that caused major incongruences most likely due to sequence saturation. Even one single branch with accelerated evolution, e.g., the branch leading to Petaluridae+Gomphidae, can affect phylogenetic hypotheses testing (see below).

Phylogeny of Anisoptera and Anisozygoptera (Epiprocta)

Divergence time estimates suggest divergence of Anisoptera and Anisozygoptera from MRCA dated from the Late Triassic ~205 Ma (fig. 2). Epiprocta as well as Anisoptera were consistent with more recent studies and recovered as monophyletic with very high support (fig. 3). We also note here that our divergence time estimates of Anisoptera tended to be younger than those found in (Letsch *et al.* 2016). The fossil-calibration approach based on penalized likelihood has been shown to overestimate true nodal age (Britton *et al.* 2007) preventing direct comparison between our dates and those estimated in (Letsch *et al.* 2016).

The problem of Gomphidae and Petaluridae

The phylogenetic position of Gomphidae and Petaluridae, both with respect to each other and the remaining anisopteran families, has long been difficult to resolve. Several phylogenetic hypothesis have been proposed in the literature based on molecular and morphological data regarding the placement of Gomphidae as sister to the remaining Anisoptera (Blanke *et al.* 2013) or to Libelluloidea (Misof *et al.* 2001). Petaluridae has exhibited stochastic relationships with

different members of Anisoptera, including sister to Gomphidae (Misof *et al.* 2001), sister to Libelluloidea (Carle *et al.* 2008), sister to Chlorogomphidae+Cordulegasteridae (Bybee *et al.* 2008) and sister to all other Anisoptera (Rehn 2003; Trueman 1996) (fig. 1C). The most resent analyses of major anisopteran lineages using several molecular markers (Carle *et al.* 2015) suggest Gomphidae and Petaluridae as a monophyletic group, but without strong branch support. Here the majority of the supermatrix analyses strongly support a sister relationship between the two families splitting from the MRCA ~165 Ma in the Middle Jurassic (fig. 2); however, almost all the coalescent-based species trees reject such a relationship with high confidence. In the presence of incomplete lineage sorting concatenation methods can be statistically inconsistent (Roch & Steel 2014) leading to an erroneous species tree topology with unreasonably high support (Kubatko & Degnan 2007). Thus, inconsistency in the recovery of a sister group relationship between Gomphidae and Petaluridae can be explained by elevated levels of incomplete lineage sorting between the families. Additionally, likelihood-mapping analysis of DNA CO and AA CO supermatrices also suggest a sister relationship between Petaluridae with Cavilabiata (fig. 3, supplementary fig. 1), whereas jackknife resampling supports a sister group relationship of Gomphidae and Petaluridae (fig. 6, supplementary fig. 1). Interestingly, our results show the branch leading to Gomphidae was found to be fast-evolving (fig. 5) and may suffer from long branch attraction thus causing a random affinity to the other groups rather than Petaluridae. Despite the fact that most of our analyses returned a sister relationship between Gomphidae and Petaluridae, further validation (i.e., by increasing taxon sampling) is required in order to make solid conclusions.

Phylogeny of Zygoptera

New zygopteran lineages originated in the Early Jurassic ~191 Ma with the early split of Lestoidea and the remaining Zygoptera (fig. 2). Subsequent occurrence of two large zygopteran groups, Calopterygoidea and Coenagrionoidea, was estimated within the Cretaceous (~145-66 Ma) and culminated with the rapid radiation of the majority of extant lineages in the Paleogene (~66-23 Ma). Our calibrated divergences generally agree with estimates in (Thomas *et al.* 2013). However, any further comparisons are virtually impossible due to the lack of comprehensive divergence time estimation for Odonata in literature.

Monophyly of Calopterygoidea

The backbone of the crown group Calopterygoidea that branched off from Coenagrionoidea ~129 Ma in the Early Cretaceous was well supported as monophyletic by various analyses (fig. 2) including jackknife site resampling of the DNA CO supermatrix (fig. 6) reaching a plateau of 100 UFBoot at 25% of the data. Previous analyses struggled to provide convincing support for the monophyly of the superfamily (Bybee *et al.* 2008; Carle *et al.* 2008; Dijkstra *et al.* 2014a) (fig. 1B); whereas only 11 out of 48 phylogenetic reconstructions rejected Calopterygoidea. Generally all analyses of AA CO and AO supermatrices exhibited decreased support for the monophyly of Calopterygoidea. Additionally, a jackknife of AA supermatrices compared to the DNA CO supermatrix demonstrated low stability of the clade possibly indicating increased sensitivity to the composition of subsampled data (fig. 6) and/or to the number of phylogenetically informative sites (Shen *et al.* 2017). This particular scenario firstly shows a lack of robustness to the data type sampled (Kumar *et al.* 2012) and secondly exemplifies statistical effects of the amount of data on stable phylogenetic inference. Theoretically, the derived estimates from a statistically consistent tree inferential method

(RoyChoudhury *et al.* 2015; Warnow 2015) applied to data without biases (systematic errors: (Yang & Rannala 2012)) converge in probability to the true value of a parameter, i.e., phylogenetic tree. In fact, our empirical results do demonstrate that with more genes added (CO 1603 vs. PO 4341 gene clusters) the inferred trees exhibit consistent recovery of monophyletic Calopterygoidea with high support. Overall both the supermatrix and supertree inferential methods using PO data, especially obtained from DNA datasets produced very similar topologies (fig. 4A) measured by pairwise normalized RF distance. However, comparison of results from the likelihood-mapping analysis of DNA CO and AA CO supermatrices shows a discrepancy where 73.9% of all quartets from DNA datasets support monophyly (fig. 3) whereas only 28.7% of AA datasets result in high support (supplementary fig. 1). In general, species trees inferred from AA data exhibit more pronounced topological differences among each other than DNA trees (fig. 5A). Topological congruency assessment of individual AA vs. DNA gene trees using RF distances indicates deeper disagreement among topologies reconstructed from AA topologies as well (fig. 5B, C) (One-sided WRST, W = 3.94e+11, $P < 2.2e-16$). Most likely these observations are attributed to the lower phylogenetic signal of individual AA gene clusters due to high sequence conservation.

Conclusions

Resolving a phylogenetic backbone for an extant group of organisms that represent a relict group, such as Odonata, has been a long standing problem in evolutionary biology, particularly among insects. Early attempts to disentangle evolutionary relationships along the backbone of Odonata left multiple questions and unresolved nodes especially within Zygoptera. Our research represents the most data rich and comprehensive analyses for phylogenetic

reconstruction for the order. We were able to provide a more detailed evolutionary picture within Zygoptera in particular identifying four stable monophyletic Calopterygoidea (fig. 2) that have been problematic for decades. Our findings within Anisoptera generally confirm previous phylogenetic re-classifications (Carle *et al.* 2015); however, here we provided more conclusive results for the long-standing problem of Petaluridae and Gomphidae relationships and identified factors that may have caused historical phylogenetic ambiguity between two families.

We also highlight the importance of every analytical step of a phylogenetic pipeline (especially where the amount and quality of data are often limited) and their impact on final tree inference starting from initial orthology detection and data processing through tree reconstruction algorithms. In the existing plethora of orthology detection methods (Kristensen *et al.* 2011) only a few are applicable for non-model organisms and often only a single method is used in phylogenetic studies (e.g., only OrthoMCL). Using three different orthology search techniques, we corroborate that more data overall provide consistent recovery of species relationships (fig. 4A, YANG50) regardless of what alignment, partition scheme or inferential method was utilized. Consistent with previous studies showing that increased character sampling leads to convergence in phylogenetic inference, sometimes converging on the wrong answer if the model of character evolution is misspecified (Gaut & Lewis 1995; Hillis *et al.* 1994; Lemmon & Moriarty 2004). Additionally, we note that all inference methods are not necessarily robust against data type (Kumar *et al.* 2012) as a general trend the majority of amino acid gene trees as well as species trees exhibited more pronounced topological differences between each other and DNA trees alike (fig. 4A, B). Trimming procedures also affected phylogenetic inference, especially in cases where no trimming or very extensive trimming was applied. For example, untrimmed fragmentary data most likely contain phylogenetic noise; thus producing

inconsistent topologies (fig. 4A), whereas trimmed data caused failure to achieve convergence within Bayesian framework. It is crucial to explore different tree building methods that take into account biological phenomena that can influence phylogenetic inference. In our case, Petaluridae and Gomphidae most likely experienced elevated levels of ILS as captured by the coalescened-based approach of ASTRAL resulting in the uncertainty of their monophyly. Also, our findings suggest that missing data contributes to taxon "roguness" and thus reduces local bootstrap support (fig. 5).

Materials and Methods

Taxon Sampling and RNA-seq

In this study, we used distinct 85 species (83 ingroup and 2 outgroup taxa). 35 RNA-seq libraries were obtained from NCBI (supplementary table 1). The remaining 58 libraries were sequenced in Bybee Lab (some species have several RNA-seq libraries). Total RNA was extracted for each taxon from eye tissue using NucleoSpin columns (Clontech) and reverse-transcribed into cDNA libraries using the Illumina TruSeq RNA v2 sample preparation kit that both generates and amplifies full-length cDNAs. Prepped mRNA libraries with insert size of ~200bp were multiplexed and sequenced on an Illumina HiSeq 2000 producing 101-bp paired-end reads by the Microarray and Genomic Analysis Core Facility at the Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT, USA. Quality scores, tissue type and other information about RNA-seq libraries are summarized in supplementary table 1.

Transcriptome Assembly and CDS Prediction

RNA-seq libraries were trimmed and de novo assembled using Trinity (Grabherr *et al.*

2011; Haas *et al.* 2013) with default parameters. Then only the longest isoform was selected from each gene for downstream analyses using Trinity utility script. In order to identify potentially coding regions within the transcriptomes, we used TransDecoder with default parameters specifying to predict only the single best ORF. Each predicted proteome was screened for contamination using DIMOND BLASTP (Buchfink *et al.* 2015) with an E-value cutoff of 10^{-10} against custom protein database. Non-arthropod hits were discarded from proteomes (amino acid, AA sequences) and corresponding CDSs. To mitigate redundancy in proteomes and CDSs, we used CD-HIT (Fu *et al.* 2012) with the identity threshold of 0.99. Such a conservative threshold was used to prevent exclusion of true paralogous sequences; thus, reducing possible false positive detection of 1:1 orthologs during homology searches.

Homology Assessment

BUSCO

BUSCO arthropod Hidden Markov Model Profiles of 2675 single-copy orthologs were used to find significant COs matches within CDS datasets by HMMER's *hmmersearch* v3 (Eddy 2011) with group-specific expected bit-score cutoffs. BUSCO classifies loci into complete [duplicated] and fragmented. Thus, only complete single-copy loci were extracted from CDS datasets and corresponding AA sequences for further phylogenetic analyses. Since loci were identified as true orthologs if they score above expected bit-score and complete if their lengths lie within ~95% of BUSCO group mean length, many partial erroneously assembled sequences were filtered out.

OrthoMCL

OrthoMCL v2.0.9 (Li *et al.* 2003) was used to compute AOs in all species using predicted AA sequences by TransDecoder. AA sequences were used in an all-vs-all BLASTP with an E-value cutoff of 10^{-10} to find putative orthologs and paralogs. The Markov Cluster algorithm (MCL) inflation point parameter was set to 2. Only 1:1 orthologs were used in further analyses. In order to exclude false-positive homology clusters identified by OrthoMCL, we applied machine learning filtering procedure (Fujimoto *et al.* 2016a) implemented in OGCleaner software v1.0 (Fujimoto *et al.* 2016b) using a metaclassifier with logistic regression.

Yang's orthology pipeline

Finally, to identify additional clusters, we used Yang's tree-based orthology inference pipeline (Yang & Smith 2014) that was specifically designed for non-model organisms using transcriptomic data. Yang's algorithm is capable of parsing paralogous gene families into "orthology" clusters that can be used in phylogenetic analyses. It has been shown that paralogous sequences encompass useful phylogenetic information (Hellmuth *et al.* 2015). First, the Transdecoder-predicted AA sequences were trimmed using CD-HIT with the identity threshold of 0.995. Then, all-vs-all BLASTP with an E-value cutoff of 10^{-5} search was implemented. The raw BLASTP output was filtered by hit fraction of 0.4. Then, MCL clustering was performed with inflation point parameter of 2. Each cluster was aligned using iterative algorithm of PASTA (Mirarab *et al.* 2015) and then was used to infer a maximum-likelihood (ML) gene tree using IQ-TREE v1.5.2 (Nguyen *et al.* 2015) with an automatic model selection. Tree tips that were longer than relative and absolute cutoffs of 0.4 and 1 respectively were removed. Mono- and paraphyletic tips that belonged to the same species were masked as well. To increase quality of homology clusters realignment, tree inference and tip masking steps were iterated with more

stringent relative and absolute masking cutoffs of 0.2 and 0.5 respectively. Finally, POs (AA sequences and corresponding CDSs) were extracted by rooted ingroups (RI) procedure using *Ephemera danica* as an outgroup (for details see (Yang & Smith 2014)).

Cluster Alignment, Trimming and Supermatrix Assembly

For most of the analyses only clusters with ≥ 42 (~50%) species present were retained. In total, we obtained five cluster types, namely DNA (CDS) and AA COs, AA AOs and DNA and AA POs. Each cluster was aligned using PASTA (Mirarab *et al.* 2015) for the DNA and AA alignments and PRANK v150803 (Loytynoja 2014; Loytynoja & Goldman 2008) for the codon alignments and alignments with removed either 1st and 2nd or 3rd codon positions. In order to reduce the amount of randomly aligned regions, we implemented ALISCORE v2.0 (Misof & Misof 2009) trimming procedure (for PASTA alignments) followed by masking any site with ≥ 42 gap characters (for both PASTA and PRANK alignments). Also, sequence fragments with >50% gap characters were removed from clusters that were subjected to ASTAL v4.10.12 (Mirarab *et al.* 2014) estimation since fragmentary data may have a negative effect on accuracy of gene and hence species tree inference (Wickett *et al.* 2014). For each of the cluster type, we assembled supermatrices from trimmed gene alignments. Additionally completely untrimmed supermatrices were generated from DNA and AA COs with ≥ 5 species present.

Phylogenetic Tree Reconstruction

Partitioning and Maximum Likelihood Inference

We inferred phylogenetic ML trees from each supermatrix using IQ-TREE implementing two partitioning schemes: single partition and those identified by PartitionFinder v2.0 (three

GTR models for DNA and a large array of protein models for AA) (Lanfear *et al.* 2016a) with relaxed hierarchical clustering option (Lanfear *et al.* 2014). In the first case, IQ-TREE was run allowing model selection and assessing nodal support with 1000 ultrafast bootstrap (UFBoot) (Minh *et al.* 2013) replicates. In the second case, IQ-TREE was run with a given PartitionFinder partition model applying gene and site resampling to minimize false-positives (Gadagkar *et al.* 2005) for 1000 UFBoot replicates.

Bayesian Inference

For Bayesian analyses implemented in ExaBayes (Aberer *et al.* 2014), we used highly trimmed (retaining sites only with occupancy of ≤ 5 gap characters) and original DNA and AA CO supermatrices assuming a single partition. We initiated 4 independent runs with 4 Markov Chain Monte Carlo (MCMC) coupled chains sampling every 500^{th} iteration. Due to high computational demands of the procedure, only the GTR and JTT substitution model priors were applied to DNA and AA CO supermatrices respectively with the default topology, rate heterogeneity and branch lengths priors. However, all supported protein substitution models as a prior was specified for the trimmed AA CO supermatrix. For convergence criteria an average standard deviation of split frequencies (ASDSF) (Lakner *et al.* 2008), a potential scale reduction factor (PSRF) (Brooks & Gelman 1998) and an effective sample size (ESS) (Lanfear *et al.* 2016b) were utilized. Values of 0% < ASDSF <1% and 1% < ASDSF < 5% indicate excellent and acceptable convergence respectively; ESS >100 and PSRF ~ 1 represent good convergence (see ExaBayes manual, (Aberer *et al.* 2014)).

Coalescent-Based Inference

ASTRAL analyses were conducted using two input types: (i) gene trees obtained by IQ-TREE allowing model selection for fully trimmed DNA and AA clusters and (ii) gene trees

obtained from the alignment-tree coestimation process in PASTA. Nodal support was assessed by local posterior probabilities (Sayyari & Mirarab 2016). ASTRAL is a statistically consistent supertree method under the multispecies coalescent model with better accuracy than other similar approaches (Mirarab *et al.* 2014).

Alignment-Free Inference

In addition to standard phylogenetic inferential approaches, we applied an alignment-free (AF) species tree estimation algorithm using Co-phylog (Yi & Jin 2013). Raw Transdecoder CDS outputs were used in this analysis using k-mer size of 9 as the half context length required for Co-phylog. Bootstrap replicate trees were obtained by running Co-Phylog with the same parameter settings on each subsampled with replacement CDS Transdecoder libraries and were used to assess nodal support.

Supernetwork reconstruction and Four-Cluster Likelihood Mapping

Using a set of 46 species trees, we constructed consensus splits networks using thresholds of 0.1 and 0.5 (majority rule) with mean edge weights in SplitsTree v4.14.4 (Huson & Bryant 2006) to visualize disagreements among individual species trees. To further investigate incongruences among species trees, we performed likelihood mapping (Strimmer & von Haeseler 1997) analyses in IQ-TREE which is analogous to the TREE-PUZZLE (Schmidt *et al.* 2002) algorithm, but computationally more efficient and allows implementation of partition models (Nguyen *et al.* 2015). Using partitioned DNA and AA supermatrices different hypotheses regarding phylogenetic relationships among certain taxonomic groups were evaluated.

Fossil Dating

A Bayesian algorithm of MCMCTREE (Yang 2007) was implemented to estimate divergence times within Odonata with 20 fossil constraints (supplementary table 3) using GTR+Γ substation model. To ensure convergence, the analysis was run independently 5 times for 10^6 generations, logging every 5th generation and then removing 10% as burn-in.

Acknowledgments

We thank Gavin Martin, Nathan Lord and Camilla Sharkey for the generation of sequence data. We also thank the Fulton Supercomputer Lab (BYU) for assistance. This work was supported by the NSF grant (SMB; DEB-1265714)

References

Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol Biol Evol* **31**, 2553-2556.

Aberer AJ, Krompass D, Stamatakis A (2013) Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology* **62**, 162-166.

Blanke A, Greve C, Mokso R, Beckmann F, Misof B (2013) An updated phylogeny of Anisoptera including formal convergence analysis of morphological characters. *Systematic Entomology* **38**, 474-490.

Breinholt JW, Earl C, Lemmon AR*, et al.* (2017) Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for Anchored Phylogenomics. *Systematic Biology*.

Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K (2007) Estimating divergence times in large phylogenetic trees. *Systematic Biology* **56**, 741-752.

Brooks SP, Gelman A (1998) General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60.

Busse S, Genet C, Hornschemeyer T (2013) Homologization of the flight musculature of zygoptera (insecta: odonata) and neoptera (insecta). *PLoS One* **8**, e55787.

Busse S, Helmker B, Hornschemeyer T (2015) The thorax morphology of Epiophlebia (Insecta: Odonata) nymphs--including remarks on ontogenesis and evolution. *Sci Rep* **5**, 12835.

Bybee S, Córdoba-Aguilar A, Duryea MC*, et al.* (2016) Odonata (dragonflies and damselflies) as a bridge between ecology and evolutionary genomics. *Front Zool* **13**, 46.

Bybee SM, Ogden TH, Branham MA, Whiting MF (2008) Molecules, morphology and fossils: a comprehensive approach to odonate phylogeny and the evolution of the odonate wing. *Cladistics-the International Journal of the Willi Hennig Society* **24**, 477-514.

Carle FL, Kjer KM, May ML (2008) Evolution of Odonata, with special reference to Coenagrionoidea (Zygoptera). *Arthropod Systematics and Phylogeny* **66**, 37-44.

Carle FL, Kjer KM, May ML (2015) A molecular phylogeny and classification of Anisoptera (Odonata). *Arthropod Systematics & Phylogeny* **73**, 281-301.

Chauhan P, Hansson B, Kraaijeveld K*, et al.* (2014) De novo transcriptome of Ischnura elegans provides insights into sensory biology, colour and vision genes. *BMC Genomics* **15**, 808.

Corbet PS (1999) *Dragonflies : behavior and ecology of Odonata* Comstock Pub. Associates, Ithaca, N.Y.

Cordoba-Aguilar A (2008) *Dragonflies and damselflies : model organisms for ecological and evolutionary research* Oxford University Press, Oxford ; New York.

Dijkstra K-DB, Bechly G, Bybee SM*, et al.* (2013) The classification and diversity of dragonflies and damselflies (Odonata). In : Zhang, Z.-Q. (Ed.) Animal Biodiversity: An Outline of Higher-level Classification and Survey of Taxonomic Richness (Addenda 2013). *Zootaxa; Vol 3703, No 1: 30 Aug. 2013DO - 10.11646/zootaxa.3703.1.9*.

Dijkstra KD, Kalkman VJ, Dow RA, Stokvis FR, Van Tol J (2014a) Redefining the damselfly families: a comprehensive molecular phylogeny of Zygoptera (Odonata). *Systematic Entomology* **39**, 68-96.

Dijkstra KD, Monaghan MT, Pauls SU (2014b) Freshwater biodiversity and aquatic insect diversification. *Annu Rev Entomol* **59**, 143-163.

Dumont HJ, Vanfleteren JR, De Jonckheere JF, PH HW (2005) Phylogenetic relationships, divergence time estimation, and global biogeographic patterns of calopterygoid damselflies (odonata, zygoptera) inferred from ribosomal DNA sequences. *Systematic Biology* **54**, 347-362.

Dumont HJ, Vierstraete A, Vanfleteren JR (2010) A molecular phylogeny of the Odonata (Insecta). *Systematic Entomology* **35**, 6-18.

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195.

Etienne RS, Rosindell J (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology* **61**, 204-213.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152.

Fujimoto MS, Suvorov A, Jensen NO, Clement MJ, Bybee SM (2016a) Detecting false positive sequence homology: a machine learning approach. *BMC Bioinformatics* **17**, 101.

Fujimoto MS, Suvorov A, Jensen NO*, et al.* (2016b) The OGCleaner: filtering false-positive homology clusters. *Bioinformatics*.

Gadagkar SR, Rosenberg MS, Kumar S (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol* **304**, 64-74.

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol Biol Evol* **12**, 152-162.

Grabherr MG, Haas BJ, Yassour M*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652.

Grimaldi DA, Engel MS (2005) *Evolution of the insects* Cambridge University Press, Cambridge, UK ; New York, NY.

Haas BJ, Papanicolaou A, Yassour M*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512.

Hasegawa E, Kasuya E (2006) Phylogenetic analysis of the insect order Odonata using 28S and 16S rDNA sequences: a comparison between data sets with different evolutionary rates. *Entomological Science* **9**, 55-66.

Hellmuth M, Wieseke N, Lechner M*, et al.* (2015) Phylogenomics with paralogs. *Proc Natl Acad Sci U S A* **112**, 2058-2063.

Hillis DM, Huelsenbeck JP, Cunningham CW (1994) Application and accuracy of molecular phylogenies. *Science* **264**, 671-677.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254-267.

Kim MJ, Jung KS, Park NS*, et al.* (2014) Molecular phylogeny of the higher taxa of Odonata (Insecta) inferred from COI, 16S rRNA, 28S rRNA, and EF1-α sequences. *Entomological Research* **44**, 65-79.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. *Brief Bioinform* **12**, 379-391.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* **56**, 17-24.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K (2012) Statistics and truth in phylogenomics. *Mol Biol Evol* **29**, 457-472.

Labandeira CC, Sepkoski JJ, Jr. (1993) Insect diversity in the fossil record. *Science* **261**, 310-315.

Lakner C, van der Mark P, Huelsenbeck JP, Larget B, Ronquist F (2008) Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology* **57**, 86-103.

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

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2016a) PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol Biol Evol*.

Lanfear R, Hua X, Warren DL (2016b) Estimating the Effective Sample Size of Tree Topologies from Bayesian Phylogenetic Analyses. *Genome Biol Evol* **8**, 2319-2332.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* **58**, 130-145.

Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. *Systematic Biology* **53**, 265-277.

Letsch H, Gottsberger B, Ware JL (2016) Not going with the flow: a comprehensive time-calibrated phylogeny of dragonflies (Anisoptera: Odonata: Insecta) provides evidence for the role of lentic habitats on diversification. *Mol Ecol* **25**, 1340-1353.

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Loytynoja A (2014) Phylogeny-aware alignment with PRANK. *Methods Mol Biol* **1079**, 155-170.

Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635.

Minh BQ, Nguyen MA, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195.

Mirarab S, Nguyen N, Guo S*, et al.* (2015) PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *J Comput Biol* **22**, 377-386.

Mirarab S, Reaz R, Bayzid MS*, et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-548.

Misof B, Liu S, Meusemann K*, et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767.

Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* **58**, 21-34.

Misof B, Rickert AM, Buckley TR, Fleck G, Sauer KP (2001) Phylogenetic signal and its decay in mitochondrial SSU and LSU rRNA gene fragments of Anisoptera. *Mol Biol Evol* **18**, 27-37.

Nee S, Holmes EC, May RM, Harvey PH (1994) Extinction rates can be estimated from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci* **344**, 77-82.

Nel A, Bechly G, Martinez-Delclos X, Fleck G (2001) A new family of Anisoptera from the Upper Jurassic of Karatau in Kazakhstan (Insecta: Odonata: Juragomphidae n. fam.). *Stuttgarter Beitraege zur Naturkunde Serie B (Geologie und Palaeontologie)* **314**, 1-9.

Nel A, Bechly G, Prokop J, Béthoux O, Fleck G (2011) Systematics and Evolution of Paleozoic And Mesozoic Damselfly-Like Odonatoptera of the 'Protozygopteran' Grade. *Journal of Paleontology* **86**, 81-104.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274.

Ogden TH, Whiting MF (2003) The problem with "the Paleoptera Problem:" sense and sensitivity. *Cladistics-the International Journal of the Willi Hennig Society* **19**, 432-442.

Pfau HK (1991) Contributions of functional morphology to the phylogenetic systematics of Odonata. *Advances in Odonatology* **5**, 109-141.

Prum RO, Berv JS, Dornburg A*, et al.* (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569-573.

Pybus OG, Harvey PH (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc Biol Sci* **267**, 2267-2272.

Rehn AC (2003) Phylogenetic analysis of higher-level relationships of Odonata. *Systematic Entomology* **28**, 181-240.

Roch S, Steel M (2014) Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol* **100C**, 56-62.

RoyChoudhury A, Willis A, Bunge J (2015) Consistency of a phylogenetic tree maximum likelihood estimator. *Journal of Statistical Planning and Inference* **161**, 73-80.

Saux C, Simon C, Spicer GS (2003) Phylogeny of the dragonfly and damselfly order odonata as inferred by mitochondrial 12S ribosomal RNA sequences. *Ann Entomol Soc Am* **96**, 693-699.

Sayyari E, Mirarab S (2016) Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Mol Biol Evol* **33**, 1654-1668.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504.

Shen XX, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol* **1**, 126.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

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

Suvorov A, Jensen NO, Sharkey CR*, et al.* (2017) Opsins have evolved under the permanent heterozygote model: insights from phylotranscriptomics of Odonata. *Mol Ecol* **26**, 1306-1322.

Thomas JA, Trueman JW, Rambaut A, Welch JJ (2013) Relaxed phylogenetics and the palaeoptera problem: resolving deep ancestral splits in the insect phylogeny. *Systematic Biology* **62**, 285-297.

Troast D, Suhling F, Jinguji H, Sahlen G, Ware J (2016) A Global Population Genetic Study of Pantala flavescens. *PLoS One* **11**, e0148949.

Trueman JWH (1996) A preliminary cladistic analysis of odonate wing venation. *Odonatologica* **25**, 59-72.

Ware J, May M, Kjer K (2007) Phylogeny of the higher Libelluloidea (Anisoptera: Odonata): an exploration of the most speciose superfamily of dragonflies. *Mol Phylogenet Evol* **45**, 289-310.

Warnow T (2015) Concatenation Analyses in the Presence of Incomplete Lineage Sorting. *PLoS Curr* **7**.

Wickett NJ, Mirarab S, Nguyen N*, et al.* (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A* **111**, E4859-4868.

Wiens JJ, Morrill MC (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic Biology* **60**, 719-731.

Yang Y, Smith SA (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol* **31**, 3081-3092.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591.

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314.

Yi H, Jin L (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res* **41**, e75.

Figures

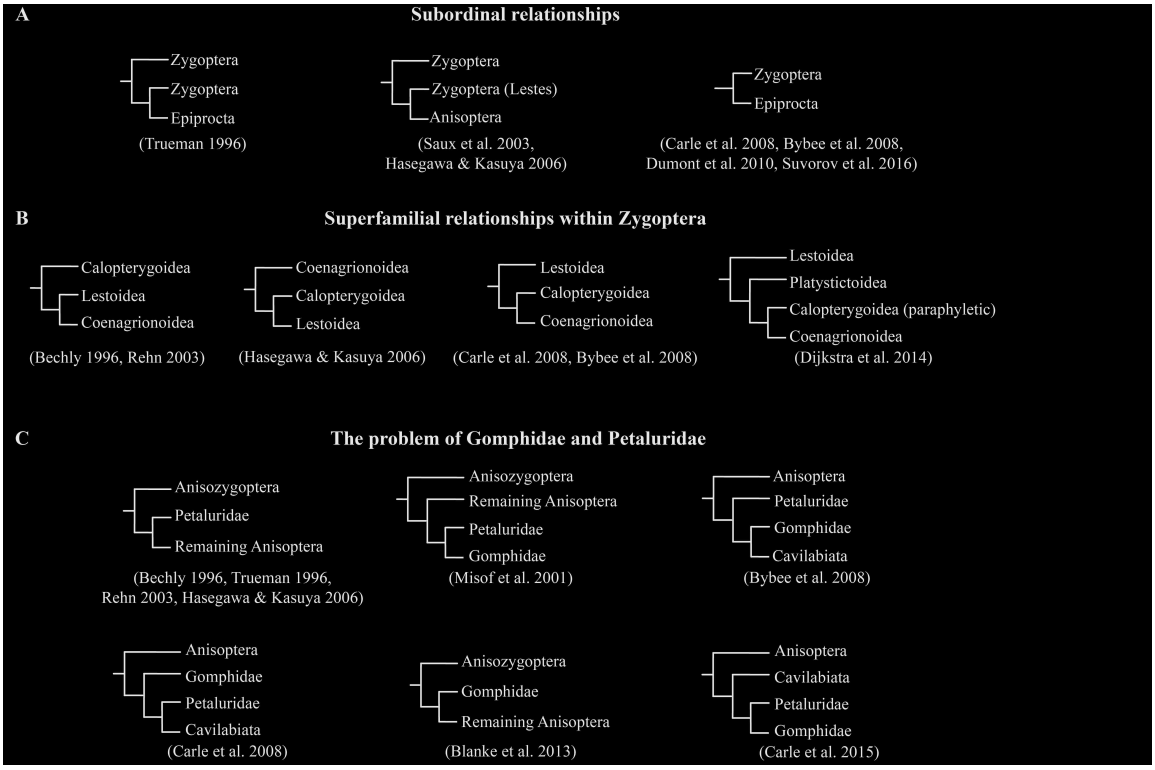


Fig. 1 Hypothesized Odonata relationships established by previous research. A. Proposed relationships between odonate Suborders. B. Hypothesized position of Calopterigoidea superfamily within Zygoptera. C. Phylogenetic relationships between Gomphidae and Petaluridae in relation to remaining Anisoptera.

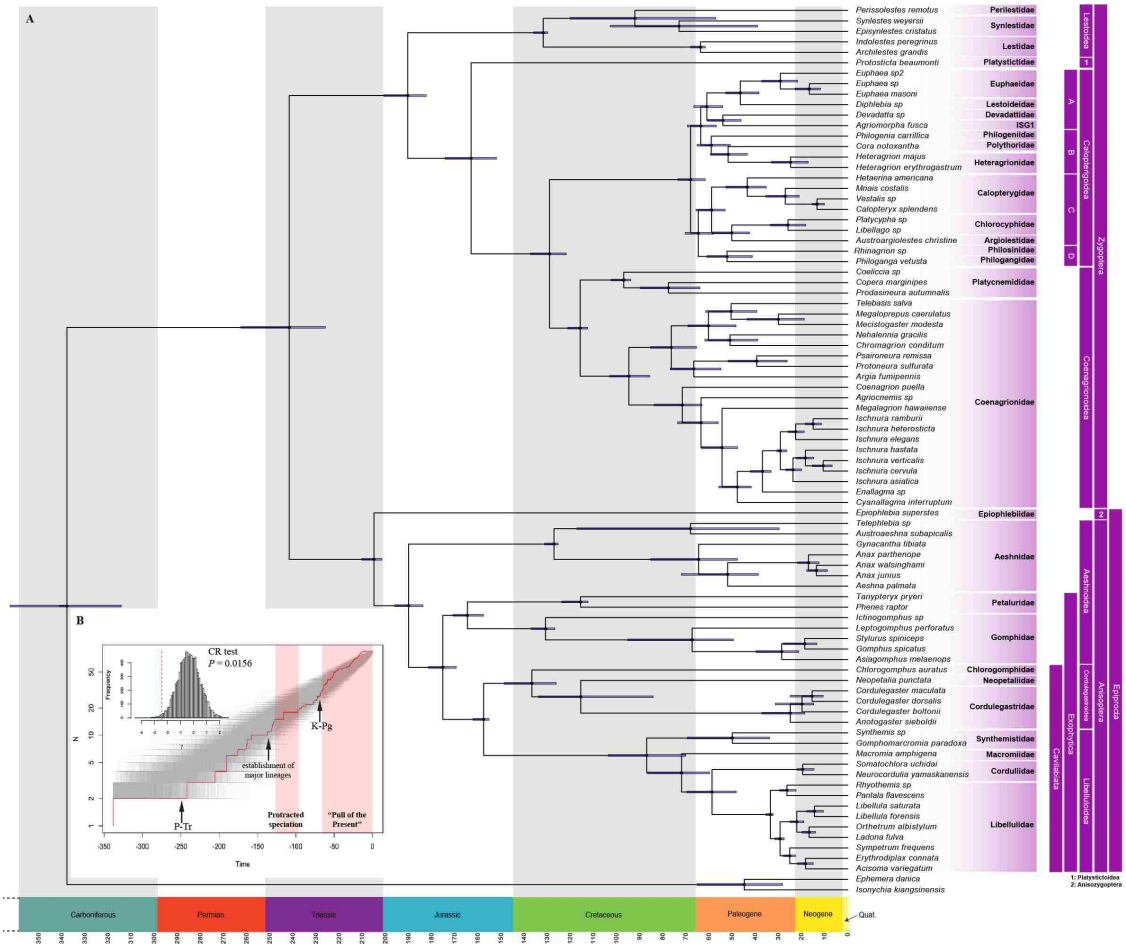


Fig. 2 Fossil calibrated ML phylogenetic tree of Odonata inferred from the CO DNA supermatrix. The white box contains Lineage Through Time Plot and null distribution of γ statistic with the dashed line representing γ statistic derived from Odonata phylogeny.

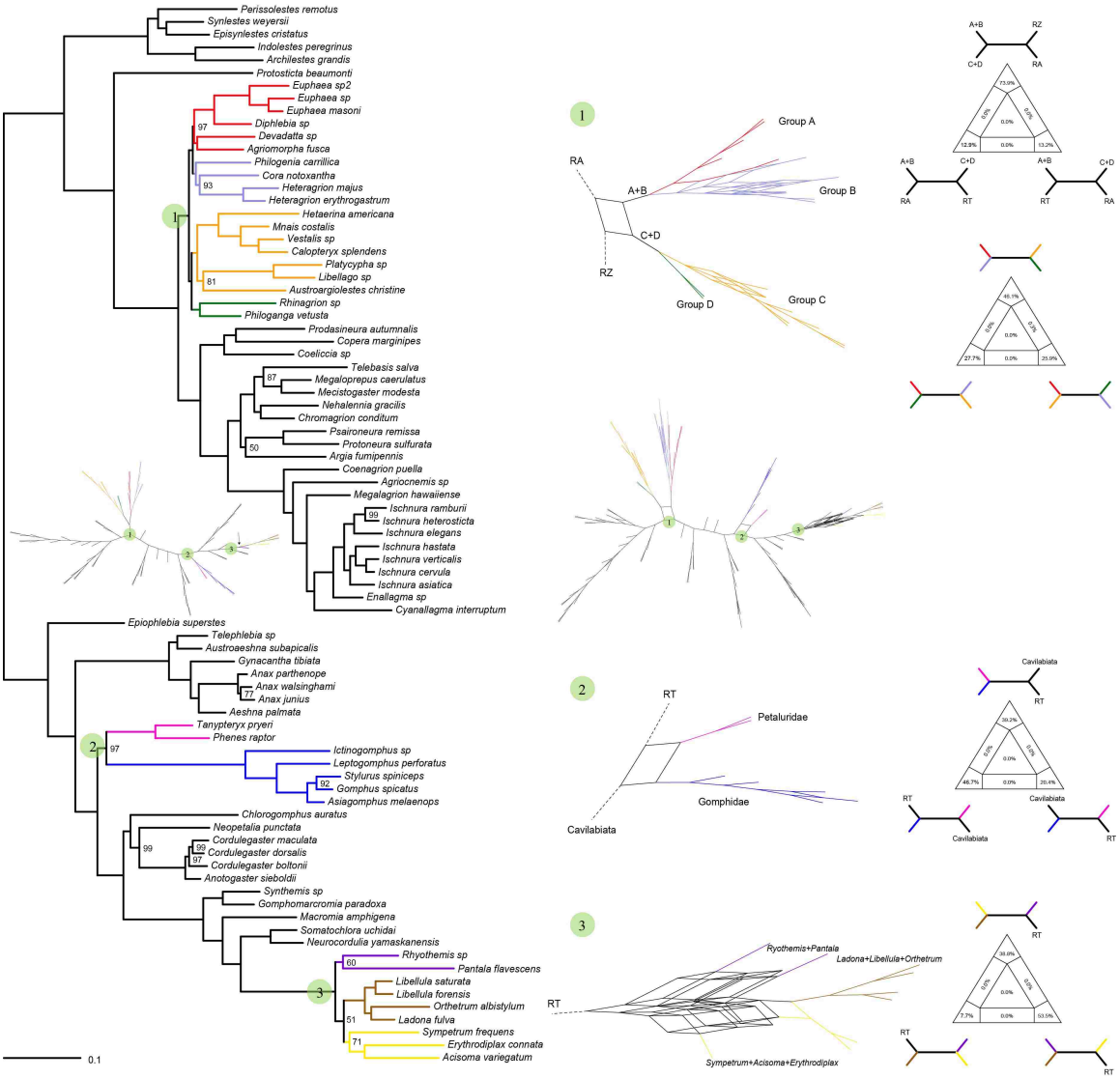


Fig. 3 ML tree inferred from the CO DNA supermatrix that used to show the most problematic topological regions (1, 2, 3). Network trees represent consensus splits networks using thresholds of 0.5 and 0.1 (left and right, respectively) with mean edge weights. The right triangles represent four-cluster maximum likelihood mapping analyses for major back-bone related phylogenetic conflicts.

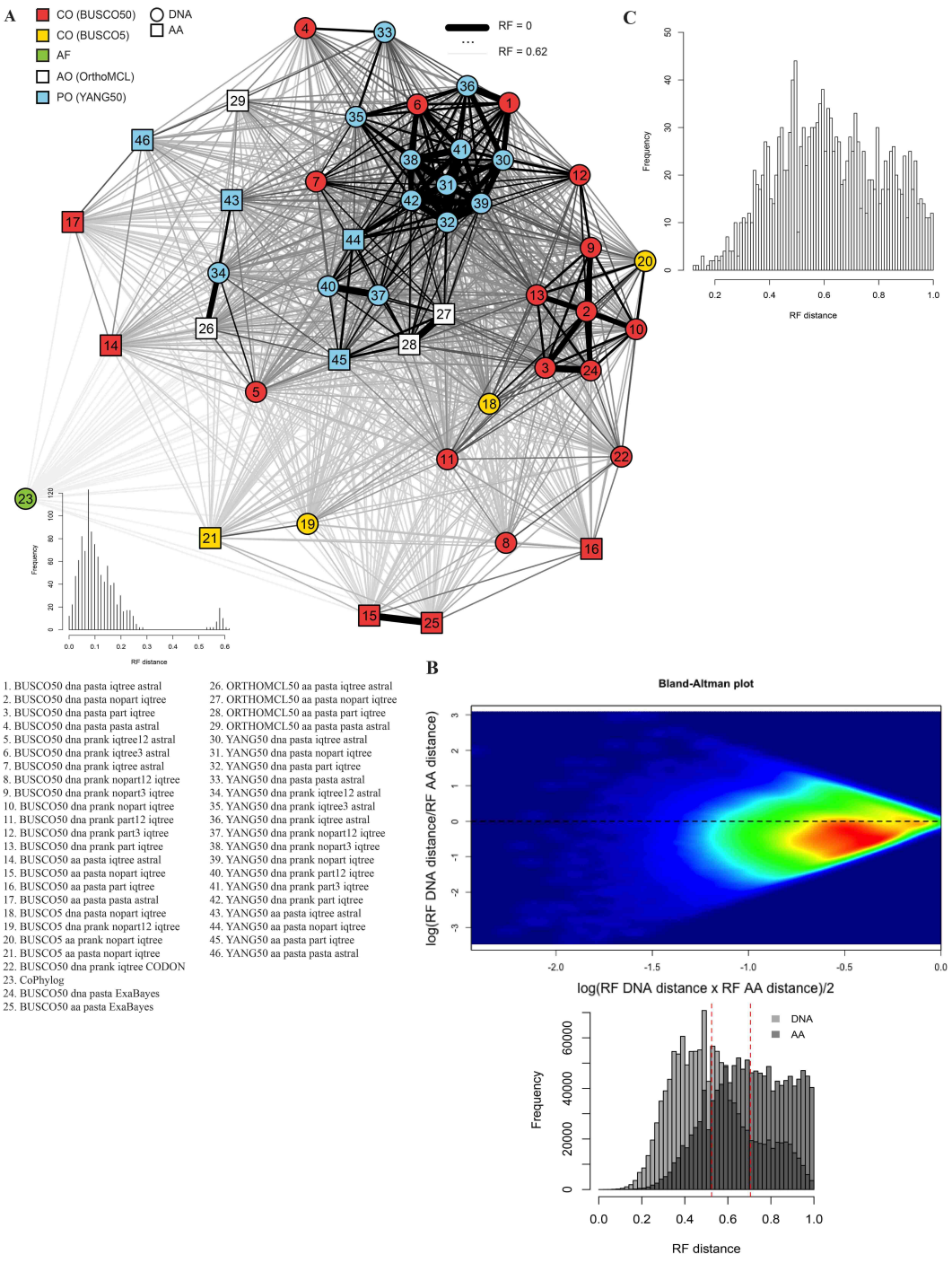


Fig. 4 A. Force-directed graph represents pairwise Robinson-Foulds (RF) distances between all phylogenetic species tree hypotheses. The histogram shows distribution of the distances. B. Bland-Altman plot that shows deeper disagreement between AA gene trees (red region) vs. DNA

gene trees. The histogram shows distribution of the RF distances among AA and DNA gene trees. C. Distribution of RF distance between gene tree topologies inferred from AA and DNA data.

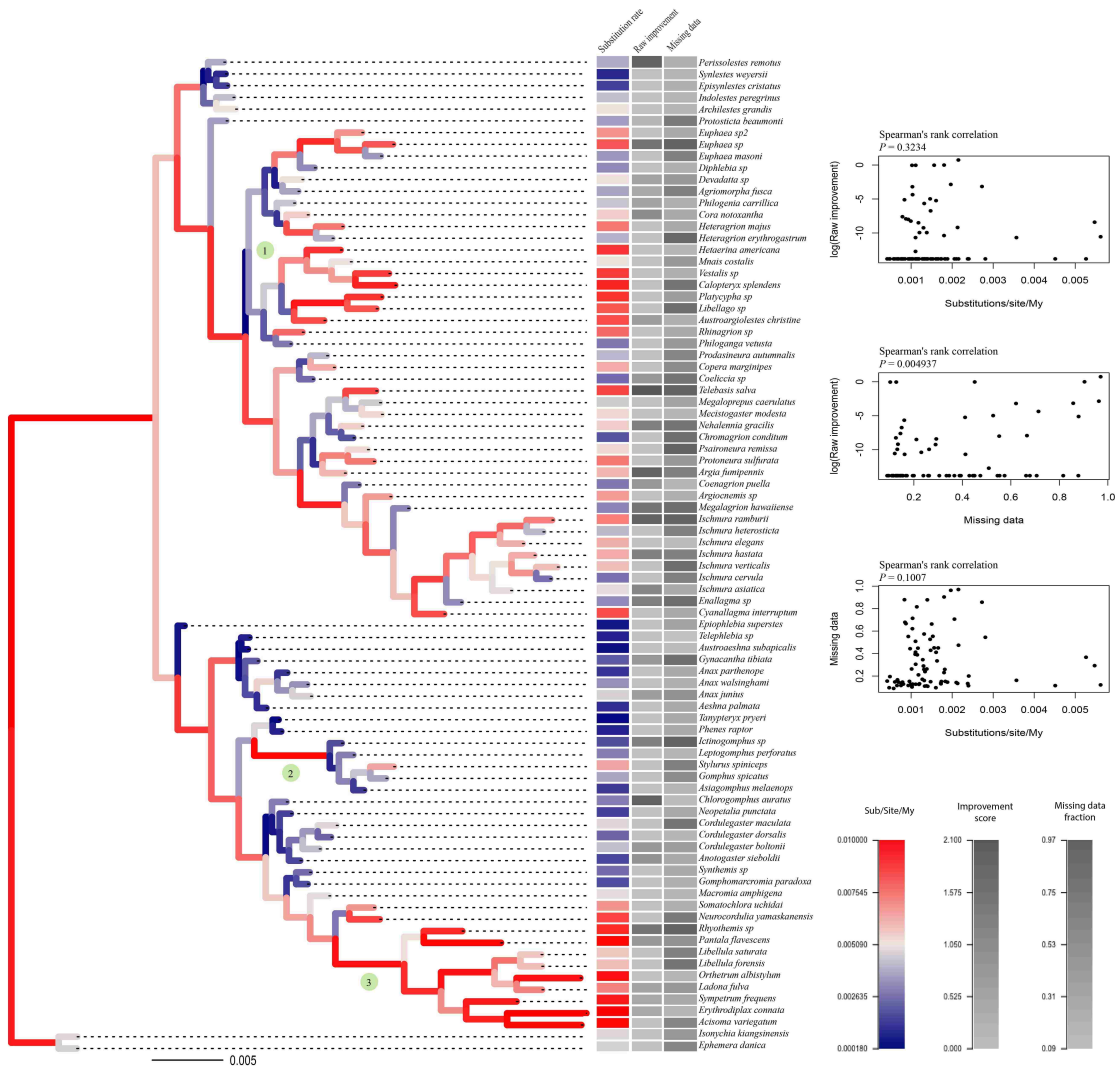


Fig. 5 ML tree inferred from the CO DNA supermatrix with branch lengths scaled to Substitution/Site/My. Red branches indicate fast evolving lineages whereas blue ones exhibit decelerated evolution. The correlation between amount of missing data, taxon "roguness" (raw improvement) inferred using RogueNaRok (Aberer *et al.* 2013) and evolutionary rates were insignificant except for missing data vs. raw improvement.

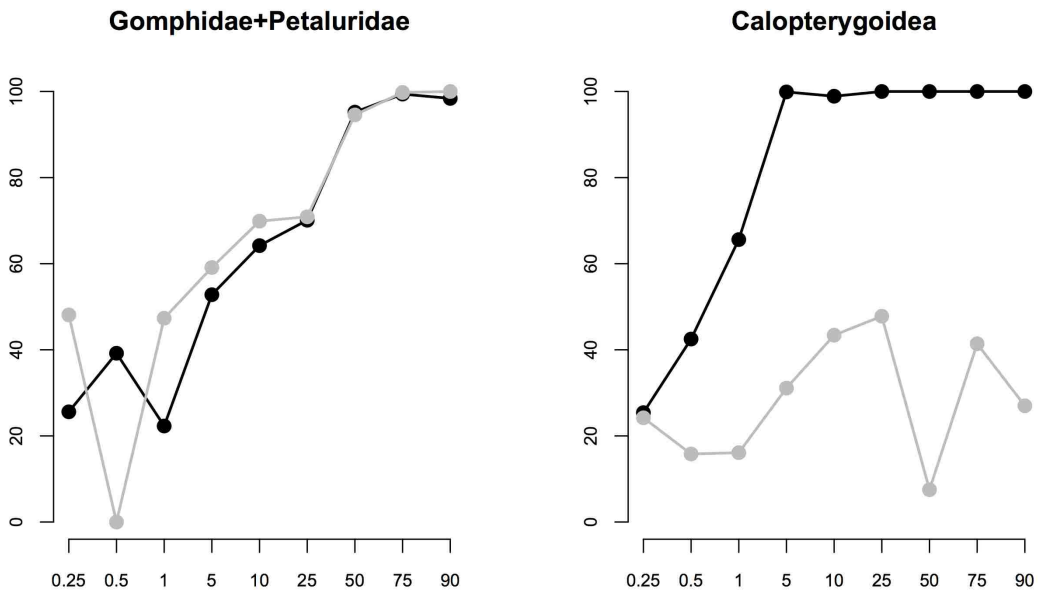


Fig. 6 Results of jackknife analyses. Black lines represent DNA data; Gray lines represent AA data. The x-axis shows percent amount of subsampled supermatrix; y-axis shows average bootstrap support from 10 replicates.

Supplementary Figures and Tables

Figures

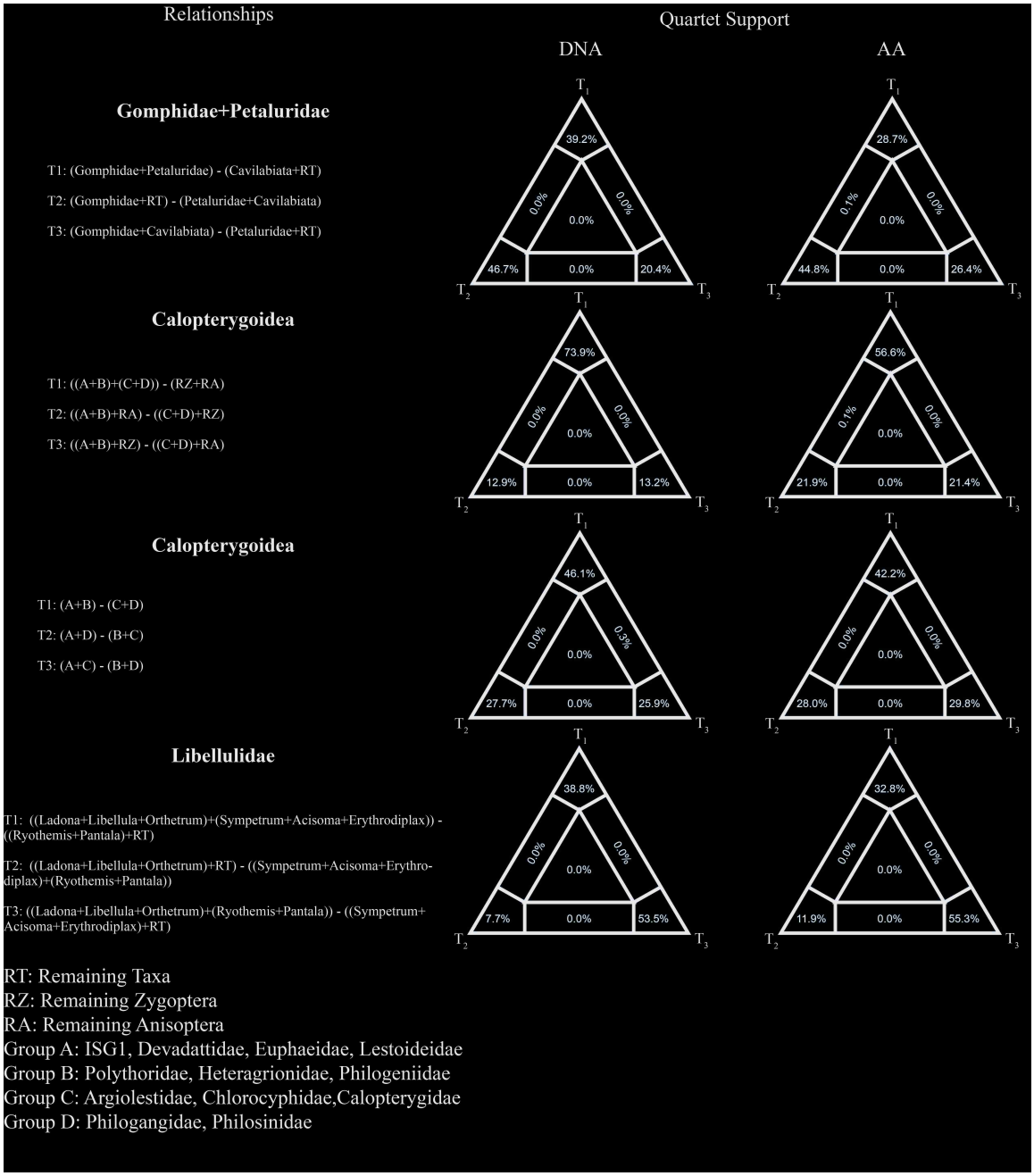


Figure S1 Four-cluster maximum likelihood mapping analyses for major back-bone related phylogenetic conflicts.

Tables

Table S1. RNA-seq read library and assembly statistics

Species	N50	Source	Platform	Body part	ID	N contigs
Acisoma variegatum	1297	Present study	HiSeq 2000	Adult Head	R_OD_267	64971
Aeshna palmata	1493	Present study	HiSeq 2000	Adult Head	AP030	50323
Agriocnemis sp.	1897	Present study	HiSeq 2000	Adult Head	R_OD_270	194794
Agriomorpha fusca	1886	Present study	HiSeq 2000	Adult Head	R_OD_214	250121
Anax junius_1	1289	Suvorov	GAIIx	Adult Head	SRR2164543	35883
Anax junius_2		Suvorov	GAIIx	Adult Head	SRR2157371	
Anax parthenope	2171	Futahashi	HiSeq 2000	Adult Head	DRR016559	49579
Anax walsinghamii	1912	Present study	HiSeq 2000	Adult Head	Anax	75979
Anotogaster sieboldii_1	2400	Futahashi	HiSeq 2000	Adult Head	DRR016590	49398
Anotogaster sieboldii_2		Futahashi	HiSeq 2000	Adult Head	DRR016589	
Archilestes grandis_1	2664	Present study	HiSeq 2000	Adult Head	AG001D	93682
Archilestes grandis_2		Present study	HiSeq 2000	Adult Head	AG001V	
Argia fumipennis violacea	920	Suvorov	GAIIx	Adult Head	SRR2157383	36769
Asiagomphus melaenops	2387	Futahashi	HiSeq 2000	Adult Head	DRR021449	64176
Austroaeshna subapicalis	1024	Present study	HiSeq 2000	Adult Head	R_OD_289	101813
Austroargiolestes christine	1671	Present study	HiSeq 2000	Adult Head	R_OD_238	132024
Calopteryx splendens	768	Misof	HiSeq 2000	Body	SRR921575	53774
Chlorogomphus auratus_1	2807	Present study	HiSeq 2000	Adult Head	R_OD388e	68737
Chlorogomphus auratus_2		Present study	HiSeq 2000	Adult Head	R_OD388m	
Chromagrion conditum	1128	Suvorov	GAIIx	Adult Head	SRR2157380	32367
Coeliccia sp.	1448	Present study	HiSeq 2000	Adult Head	R_OD_219	51527
Coenagrion puella	1800	Johnston	HiSeq 2000	Body (18 pooled)	SRR740826	111962
Copera manginipes	1443	Present study	HiSeq 2000	Adult Head	R_OD205	60985
Cordulegaster boltonii	1261	Misof	HiSeq 2000	Body	SRR921583	72933
Cordulegaster dorsalis	2450	Present study	HiSeq 2000	Adult Head	R_OD390e	50164
Cordulegaster maculata	895	Suvorov	GAIIx	Adult Head	SRR2164542	29392
Cyanallagma interruptum	1824	Present study	HiSeq 2000	Adult Head	R_OD380	103460
Devedatta sp.	1773	Present study	HiSeq 2000	Adult Head	R_OD386e	54507
Diphlabia sp.	2118	Present study	HiSeq 2000	Adult Head	R_OD_295	93103

Enallagma sp	876	Suvorov	GAIIx	Adult Head	SRR2157367	29457
Epiophlebia superstes	1408	Misof	HiSeq 2000	Body	SRR921592	75483
Episylestes cristatus	2007	Present study	HiSeq 2000	Adult Head	R_OD_304	185000
Erythrodiplax conata	2767	Present study	HiSeq 2000	Adult Head	R_OD371	76443
Euphaea masoni	1022	Present study	HiSeq 2000	Adult Head	R_OD228	71683
Euphaea sp.	1843	Present study	HiSeq 2000	Adult Head	R_OD322	83118
Euphaea sp.2	386	Present study	HiSeq 2000	Adult Head	R_OD_086	18371
Gomphomarcromia paradoxa	2139	Present study	HiSeq 2000	Adult Head	R_OD374	82401
Gomphus spicatus	1266	Suvorov	GAIIx	Adult Head	SRR2157378	38919
Gynacantha tibiata_1	560	Present study	HiSeq 2000	Adult Head	R_OD_107E	79925
Gynacantha tibiata_2		Present study	HiSeq 2000	Adult Head	R_OD_107M	
Hetaerina americana_1	2461	Present study	HiSeq 2000	Adult Head	HA001D	116668
Hetaerina americana_2		Present study	HiSeq 2000	Adult Head	HA001V	
Heteragrion erythrogastrum	401	Present study	HiSeq 2000	Adult Head	R_OD_133E	80468
Heteragrion majus	1612	Present study	HiSeq 2000	Adult Head	R_OD_137	80316
Ictinogomphus_sp	488	Present study	HiSeq 2000	Adult Head	R_OD_088	29155
Indolestes peregrinus_1	2909	Futahashi	HiSeq 2000	Adult Head (dorsal eyes)	DRR022343	76455
Indolestes peregrinus_2		Futahashi	HiSeq 2000	Adult Head (ventral eyes)	DRR022348	
Indolestes peregrinus_3		Futahashi	HiSeq 2000	Larval Head	DRR016542	
Ischnura asiatica	2317	Futahashi	HiSeq 2000	Adult Head	DRR021460	67337
Ischnura cervula	1003	Suvorov	GAIIx	Adult Head	SRR2157372	43487
Ischnura elegans	1713	Chauhan	HiSeq 2000	Body	SRR1265958	139682
Ischnura hastata	1493	Suvorov	GAIIx	Adult Head	SRR2164548	43693
Ischnura heterosticta_1	487	Present study	HiSeq 2000	Adult Head	R_OD_081	194456
Ischnura heterosticta_2		Present study	HiSeq 2000	Adult Head	R_OD_081	
Ischnura ramburii	742	Speiser	454	Adult Head	SRR2103464	14370
Ischnura verticalis	796	Suvorov	GAIIx	Adult Head	SRR2164552	28678
Ladona fulva	1632	LF_GENOME	HiSeq 2000	Body	SRR1850403	74562
Leptogomphus perforatus_1	2246	Present study	HiSeq 2000	Adult Head	R_OD_224	94772
Leptogomphus perforatus_2	2246	Present study	HiSeq 2000	Adult Head	R_OD_225	
Libellago_sp_1	404	Present study	HiSeq 2000	Adult Head	R_OD_084	125966
Libellago_sp_2		Present study	HiSeq 2000	Adult Head	R_OD_084	

Libellago_sp_3		Present study	HiSeq 2000	Adult Head	R_OD_085	
Libellula forensis	1044	Suvorov	GAIIx	Adult Head	SRR2164546	33141
Libellula saturnata	1253	Suvorov	GAIIx	Adult Head	SRR2164547	36735
Macromia amphigena	2690	Futahashi	HiSeq 2000	Adult Head	DRR021501	60000
Mecistogaster modesta	1845	Present study	HiSeq 2000	Adult Head	Sample_OD_110	102272
Megalagrion hawaiiensis_1	477	Present study	HiSeq 2000	Adult Head	R_OD_080	115410
Megalagrion hawaiiensis_2		Present study	HiSeq 2000	Adult Head	R_OD_080	
Megaloprepus caerulatus_1	1942	Present study	HiSeq 2000	Adult Head	R_OD111	86551
Megaloprepus caerulatus_2		Present study	HiSeq 2000	Adult Head	R_OD111e	
Mnais costalis	1756	Futahashi	HiSeq 2000	Adult Head	DRR021499	41735
Nehalennia gracilis	909	Suvorov	GAIIx	Adult Head	SRR2157379	35044
Neopetalia punctata	1964	Present study	HiSeq 2000	Adult Head	R_OD376	106056
Neurocordulia yamaskanensis	1129	Suvorov	GAIIx	Adult Head	SRR2157382	36412
Orthetrum albistylum_1	3875	Futahashi	HiSeq 2000	Adult Head (dorsal eyes)	DRR016166	132809
Orthetrum albistylum_2		Futahashi	HiSeq 2000	Adult Head (ventral eyes)	DRR016167	
Orthetrum albistylum_3		Futahashi	HiSeq 2000	Larval Head	DRR016170	
Pantala flavscens	1514	Nankai Uni	HiSeq 2000	Body	SRR1184263	44231
Perrisolestes remotus	1894	Present study	HiSeq 2000	Adult Head	R_OD123	86873
Phenes raptor	2050	Present study	HiSeq 2000	Adult Head	R_OD378	85039
Philoganga vetusta	1956	Present study	HiSeq 2000	Adult Head	R_OD384e	55597
Philogenia carillaca	2223	Present study	HiSeq 2000	Adult Head	R_ OD_135	80247
Platychypha sp.	1573	Present study	HiSeq 2000	Adult Head	R_OD_253	65870
Cora notoxantha	1740	Present study	HiSeq 2000	Adult Head	Sample_OD_136	89059
Prodasineura automalis	634	Present study	HiSeq 2000	Adult Head	R_ OD_230	162702
Protoneura sulfurata	1408	Present study	HiSeq 2000	Adult Head	R_OD128	64844
Protosticta beaumonti	1245	Present study	HiSeq 2000	Adult Head	R_OD211	53445
Psaironeura remissa	344	Present study	HiSeq 2000	Adult Head	R_OD_121E	114820
Rhinoagrion sp.	1677	Present study	HiSeq 2000	Adult Head	R_OD348	86872
Rhyothemis sp.	513	Present study	HiSeq 2000	Adult Head	R_OD_087	38391
Somatochlora uchidai	2405	Futahashi	HiSeq 2000	Adult Head	DRR016591	53146
Stylurus spiniceps	1118	Suvorov	GAIIx	Adult Head	SRR2157381	38792
Sympetrum frequens_1	2257	Futahashi	HiSeq 2000	Adult Head (dorsal	DRR016603	72617

				eyes)		
Sympetrum frequens_2		Futahashi	HiSeq 2000	Adult Head (ventral eyes)	DRR016604	
Sympetrum frequens_3		Futahashi	HiSeq 2000	Larval Head	DRR022958	
Synlestes weyersii	1512	Present study	HiSeq 2000	Adult Head	R_ OD_248	134948
Synthemis sp.	1773	Present study	HiSeq 2000	Adult Head	R_OD305	74552
Tanypteryx pryeri_1	2067	Futahashi	HiSeq 2000	Adult Head (dorsal eyes)	DRR022969	79789
Tanypteryx pryeri_2		Futahashi	HiSeq 2000	Adult Head (ventral eyes)	DRR022970	
Telebasis salva	601	Speiser	454	Adult Head	SRR2103540	9273
Telephlebia sp.	2261	Present study	HiSeq 2000	Adult Head	R_OD_249	152885
Vestalis sp.	1955	Present study	HiSeq 2000	Adult Head	R_OD082	69090
Outgroup						
Ephemera danica	1035	BCM	HiSeq 2000	Body	SRR1793079	131807
Isonychia kiangsinensis	1027	Nankai Uni	HiSeq 2000	Body	SRR1184396	105384

_1,_2: biological RNA-seq
library replicates
Suvorov: Suvorov et al. (2017) Opsins have evolved under the
permanent heterozygote model: insights from phylotranscriptomics
of Odonata. Mol Ecol
Futahashi: Futahashi et al. (2015) Extraordinary
diversity of visual opsin genes in dragonflies.
Proc Natl Acad Sci U S A
Speiser: Speiser (2014) Using phylogenetically-informed annotation (PIA) to search for light-interacting genes in
transcriptomes from non-model organisms. BMC bioinformatics
Chauhan: Chauhan et al. (2014) De novo transcriptome of Ischnura
elegans provides insights into sensory biology, colour and vision
genes. BMC Genomics
Misof: Misof et al. (2014) Phylogenomics
resolves the timing and pattern of insect
evolution. Science
Johnston: Johnston and Rolff (2013) Immune- and wound-
dependent differential gene expression in an ancient insect. Dev
Comp Immunol
BCM, Nankai Uni: Unpublished
(only NCBI records)

Table S2. Supermatrix statistics

Tree name	Input	Gene_cluster_type	PASTA	PRANK	Supermatrix	PartitionFinder	IQTREE	ASTRAL	ALISCORE	50%> occupancy site masking	fragments >50% gaps removed
BUSCO50_dna_pasta_nopart_iqtree.tre	BUSCO50 DNA (1603 gene loci)	CO	x	NA	x	NA	x	NA	x	x	NA
BUSCO50_dna_pasta_part_iqtree.tre	BUSCO50 DNA (1603 gene loci)	CO	x	NA	x	x	x	NA	x	x	NA
50BUSCO_dna_prank_nopart12_iqtree.tre	BUSCO50 DNA (1603 gene loci) 1st2nd	CO	NA	x	x	NA	x	NA	NA	x	NA
50BUSCO_dna_prank_part12_iqtree.tre	BUSCO50 DNA (1603 gene loci) 1st2nd	CO	NA	x	x	x	x	NA	NA	x	NA
50BUSCO_dna_prank_nopart_iqtree.tre	BUSCO50 DNA (1603 gene loci)	CO	NA	x	x	NA	x	NA	NA	x	NA
50BUSCO_dna_prank_part_iqtree.tre	BUSCO50 DNA (1603 gene loci)	CO	NA	x	x	x	x	NA	NA	x	NA
BUSCO50_dna_pasta_pasta_astral.tre	BUSCO50 DNA (1603 gene loci)	CO	x	NA	NA	NA	NA	x	NA	NA	NA
BUSCO50_dna_pasta_iqtree_astral.tre	BUSCO50 DNA (1603 gene loci)	CO	x	NA	NA	NA	x	x	x	x	x
BUSCO50_dna_prank_iqtree_astral.tre	BUSCO50 DNA (1603 gene loci)	CO	NA	x	NA	NA	x	x	NA	x	x
BUSCO50_dna_prank_iqtree12_astral.tre	BUSCO50 DNA (1603 gene loci) 1st2nd	CO	NA	x	NA	NA	x	x	NA	x	x
BUSCO50_dna_prank_iqtree3_astral.tre	BUSCO50 DNA (1603 gene loci) 3rd	CO	NA	x	NA	NA	x	x	NA	x	x
BUSCO50_dna_prank_part3_iqtree.tre	BUSCO50 DNA (1603 gene loci) 3rd	CO	x	NA	x	x	x	NA	x	x	NA
BUSCO50_dna_prank_nopart3_iqtree.tre	BUSCO50 DNA (1603 gene loci) 3rd	CO	NA	x	x	NA	x	NA	NA	x	NA
BUSCO50_dna_pasta_nopart_e	BUSCO50 DNA	CO	x	NA	x	NA	NA	NA	x	x	NA

xabayes.tre	(1603 gene loci)										
BUSCO50_dna_pasta_nopart_reduced_exabayes.tre	BUSCO50 DNA (1603 gene loci) trimmed	CO	x	NA	x	NA	NA	NA	x	x (90%)	NA
CODON_ECMK07_GY1KTS_BUSCO50_dna_prank_iqtree.tre	BUSCO50 DNA (1603 gene loci)	CO	NA	x	x	NA	x	NA	NA	x	NA
BUSCO5ormore_dna_pasta_nopart_iqtree.tre	BUSCOALL DNA (2449 gene loci)	CO	x	NA	x	NA	x	NA	NA	NA	NA
BUSCO5ormore_dna_prank_nopart_iqtree.tre	BUSCOALL DNA (2449 gene loci)	CO	NA	x	x	NA	x	NA	NA	NA	NA
BUSCO5ormore_dna_prank_nopart12_iqtree.tre	BUSCOALL DNA (2449 gene loci)1st2nd	CO	NA	x	x	NA	x	NA	NA	NA	NA
BUSCO50_prot_pasta_nopart_iqtree.tre	BUSCO50 AA (1603 gene loci)	CO	x	NA	x	NA	x	NA	x	x	NA
BUSCO50_prot_pasta_part_iqtree.tre	BUSCO50 AA (1603 gene loci)	CO	x	NA	x	x	x	NA	x	x	NA
BUSCO50_prot_pasta_pasta_astral.tre	BUSCO50 AA (1603 gene loci)	CO	x	NA	NA	NA	NA	x	NA	NA	NA
BUSCO50_prot_pasta_iqtree_astral.tre	BUSCO50 AA (1603 gene loci)	CO	x	NA	NA	NA	x	x	x	x	x
BUSCO50_dna_pasta_nopart_exabayes.tre	BUSCO50 AA (1603 gene loci)	CO	x	NA	x	NA	NA	NA	x	x	NA
BUSCO50_dna_pasta_nopart_reduced_exabayes.tre	BUSCO50 AA (1603 gene loci) trimmed	CO	x	NA	x	NA	NA	NA	x	x (90%)	NA
BUSCO5ormore_prot_pasta_nopart_iqtree.tre	BUSCOALL AA (2449 gene loci)	CO	x	NA	x	NA		NA	NA	NA	NA
ORTHOMCL50_prot_pasta_iqtree_astral.tre	OrthoMCL AA (1643 gene loci)	AO	x	NA	x	NA	x	NA	x	x	NA
ORTHOMCL50_prot_pasta_nopart_iqtree.tre	OrthoMCL AA (1643 gene loci)	AO	x	NA	x	x	x	NA	x	x	NA

ORTHOMCL50_prot_pasta_part_iqtree.tre	OrthoMCL AA (1643 gene loci)	AO	x	NA	NA	NA	NA	x	NA	NA	NA
ORTHOMCL50_prot_pasta_pasta_astral.tre	OrthoMCL AA (1643 gene loci)	AO	x	NA	NA	NA	x	x	x	x	x
YANG50_dna_pasta_nopart_iqtree.tre	YANG50 DNA (4341 gene loci)	PO	x	NA	x	NA	x	NA	x	x	NA
YANG50_dna_pasta_part_iqtree.tre	YANG50 DNA (4341 gene loci)	PO	x	NA	x	x	x	NA	x	x	NA
YANG50_dna_prank_nopart12_iqtree.tre	YANG50 DNA (4341 gene loci) 1st2nd	PO	NA	x	x	NA	x	NA	NA	x	NA
YANG50_dna_prank_part12_iqtree.tre	YANG50 DNA (4341 gene loci) 1st2nd	PO	NA	x	x	x	x	NA	NA	x	NA
YANG50_dna_prank_nopart_iqtree.tre	YANG50 DNA (4341 gene loci)	PO	NA	x	x	NA	x	NA	NA	x	NA
YANG50_dna_prank_part_iqtree.tre	YANG50 DNA (4341 gene loci)	PO	NA	x	x	x	x	NA	NA	x	NA
YANG50_dna_pasta_pasta_astral.tre	YANG50 DNA (4341 gene loci)	PO	x	NA	NA	NA	NA	x	NA	NA	NA
YANG50_dna_pasta_iqtree_astral.tre	YANG50 DNA (4341 gene loci)	PO	x	NA	NA	NA	x	x	x	x	x
YANG50_dna_prank_iqtree_astral.tre	YANG50 DNA (4341 gene loci)	PO	NA	x	NA	NA	x	x	NA	x	x
YANG50_dna_prank_iqtree12_astral.tre	YANG50 DNA (4341 gene loci) 1st2nd	PO	NA	x	NA	NA	x	x	NA	x	x
YANG50_dna_prank_iqtree3_astral.tre	YANG50 DNA (4341 gene loci) 3rd	PO	NA	x	NA	NA	x	x	NA	x	x
YANG50_dna_prank_part3_iqtree.tre	YANG50 DNA (4341 gene loci) 3rd	PO	x	NA	x	x	x	NA	x	x	NA
YANG50_dna_prank_nopart3_iqtree.tre	YANG50 DNA (4341 gene loci) 3rd	PO	NA	x	x	NA	x	NA	NA	x	NA

YANG50_prot_pasta_nopart_iqtree.tre	YANG50 AA (4290 gene loci)	PO	x	NA	x	NA	x	NA	x	x	NA
YANG50_dna_pasta_part_iqtree.tre	YANG50 AA (4290 gene loci)	PO	x	NA	x	x	x	NA	x	x	NA
YANG50_prot_pasta_pasta_astral.tre	YANG50 AA (4290 gene loci)	PO	x	NA	NA	NA	NA	x	NA	NA	NA
YANG50_prot_pasta_iqtree_astral.tre	YANG50 AA (4290 gene loci)	PO	x	NA	NA	NA	x	x	x	x	x
CoPhylog.tre	CoPhylog	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

CO: conservative single-copy orthologs (BUSCO)
AO: all single-copy orthologs (OrthoMCL)
PO: paralogy-parsed orthologs (Young)

Table S3. Fossil calibration points, age constraints and priors used for estimation of divergence times in MCMCTREE.

Fossil Taxon	Lineage	Node #	Minimum Age (Ma)	Maximum Age (Ma)	First Appearance Confidence Interval	Skewed Normal Prior
Severinula leopoldi	Palaeoptera	1	314.6	363.29	Continious sampling	Unif(314.6, 363.29)
Triassolestodes asiaticus	Odonata	2	221.5	300	NA	Unif(221.5, 300)
Liassophlebia sp.	Epiprocta	3	201.6	214.93	Oldest gap	SN(201.6, 3.5, 50)
Henrotayia marci	Anisoptera	4	183	220.17	Oldest gap	SN(183, 12, 50)
Palaeaeschna vidali	Aeshnidae	5	125.5	142.04	Continious sampling	SN(125.5, 5, 50)
Jibeigomphus xinboensis	Gomphidae	6	130	164.7	Bounded by Exophytica	Unif(130, 164.7)
Croatocordulia platyptera	Corduliidae	7	11.61	21.61	NA	Unif(11.61, 21.61)
Tauriphila? cerestensis	Libellulidae	8	23.03	33.03	NA	Unif(23.03, 33.03)
Eosagrion risi	Zygoptera	9	182	204.83	Percentiles of gap sizes	SN(182, 7, 50)
Juralibellula ningchengensis	Cavilabiata	10	155.7	164.4	Bounded by Exophytica	Unif(155.7,164.4)
Cretalestes martinae	Lestoidea	11	130	143.09	Continious sampling	SN(130, 4, 50)
Lestes conexus	Lestidae	12	61.7	72.32	Continious sampling	SN(61.7, 3, 50)
Sinocalopteryx shangyongensis	Calopterygoidea	13	50.3	63.32	Continious sampling	SN(50.3, 4, 50)
Ischnura velteni	Ischnura	14	20.43	30.43	NA	Unif(20.43, 30.43)
Palaeodisparoneura burmanica	Platycnemididae	15	94.3	104.3	NA	Unif(94.3, 104.3)
Calopteryx andancensis	Calopteryx	16	5.3	15.3	NA	Unif(5.3, 15.3)
Argentinopetala archangelskyi	Petaluridae	17	112.6	142.6	NA	ST(112.6, 5, 20, 3)
Mesochlorogomphus crabbi	Chlorogomphidae	18	125.5	155.5	NA	ST(125.5, 5, 20, 3)
Eoprotoneura hyperstigma	Coenagrionoidea	19	112.6	145.23	Continious sampling	SN(112.6, 10, 50)
Pheugothemis westwoodi	Exophytica	20	164.7	183	Bounded by Anisoptera	Unif(164.7, 183)

Manuscript published in BMC Bioinformatics

Chapter 3

Detecting false positive sequence homology: a machine learning approach

M. Stanley Fujimoto[1*], Anton Suvorov[2*§], Nicholas O. Jensen[2*], Mark J. Clement[1], and Seth M. Bybee[2]

[1]Computer Science Department, Brigham Young University, Provo, Utah 84602

[2]Department of Biology, Brigham Young University, Provo, Utah 84602

[*] Equal contributions

[§] Corresponding author

Abstract

Accurate detection of homologous relationships of biological sequences (DNA or amino acid) amongst organisms is an important and often difficult task that is essential to various evolutionary studies, ranging from building phylogenies to predicting functional gene annotations. There are many existing heuristic tools, most commonly based on bidirectional BLAST searches that are used to identify homologous genes and combine them into two fundamentally distinct classes: orthologs and paralogs. Due to only using heuristic filtering based on significance score cutoffs and having no cluster post-processing tools available, these methods can often produce multiple clusters constituting unrelated (non-homologous) sequences. Therefore sequencing data extracted from incomplete genome/transcriptome assemblies

originated from low coverage sequencing or produced by *de novo* processes without a reference genome are susceptible to high false positive rates of homology detection.

Results

In this paper we develop biologically informative features that can be extracted from multiple sequence alignments of putative homologous genes (orthologs and paralogs) and further utilized in context of guided experimentation to verify false positive outcomes. We demonstrate that our machine learning method trained on both known homology clusters obtained from OrthoDB and randomly generated sequence alignments (non-homologs), successfully determines apparent false positives inferred by heuristic algorithms especially among proteomes recovered from low-coverage RNA-seq data. Almost ~42% and ~25% of predicted putative homologies by InParanoid and HaMStR respectively were classified as false positives on experimental data set.

Conclusions

Our process increases the quality of output from other clustering algorithms by providing a novel post-processing method that is both fast and efficient at removing low quality clusters of putative homologous genes recovered by heuristic-based approaches.

Background

One of the most fundamental questions of modern comparative evolutionary phylogenomics is to identify common (homologous) genes that originated through complex biological mechanisms such as speciation, multiple gene losses/gains, horizontal gene transfers, deep coalescence, etc. (Koonin 2005). When homologous sequences are identified, they are usually grouped and aligned together to form clusters. Homologous DNA (and those translated to amino acids) sequences can be further subdivided into two major classes: orthologs and paralogs. Orthologs are defined as homologous genes in different species that arose due to speciation events, whereas paralogs have evolved from gene duplications. Moreover, orthologous genes are more likely to exhibit a similar tempo and mode of evolution, thus preserving overall sequence composition and physiological function. Paralogs, instead, tend to follow different evolutionary trajectories leading to subfunctionalization, neofunctionalization or both (Gabaldon & Koonin 2013). Nevertheless this phenomenon, called the ortholog conjecture, is still debatable (Dessimoz *et al.* 2012) and requires additional validation since it has been shown that even between closely related species some orthologs can diverge such that they eventually loose common functionality.

The accurate detection of sequence homology and subsequent binning into aforementioned classes is essential for robust reconstruction of evolutionary histories in the form of phylogenetic trees (Delsuc *et al.* 2005). To date, numerous computational algorithms and statistical methods have been developed to perform orthology/paralogy assignments for genic sequences (for review see (Kristensen *et al.* 2011)). Methodologically these approaches employ heuristic-based or evidence (phylogenetic tree)-based identification strategies, which produces varying frequencies of false positive or negative results. The majority of heuristic algorithms rely

on the principle of Reciprocal Best Hit (RBH, (Overbeek *et al.* 1999)) where BLAST (Altschul *et al.* 1990) hit scores (e-values) approximate evolutionary similarity between two biological sequences. Further algorithmic augmentations of those heuristics, for instance Markov graph clustering (unsupervised learning) (Li *et al.* 2003), enables the definition of orthologous/paralogous clusters from multiple pairwise comparisons. Despite their relatively low computational complexity, these algorithms have been shown to overestimate the number of putative homologies (i.e., higher rates of false positive detection compared to evidence-based methods (Chen *et al.* 2007)).

In this current era of next-generation sequence data researchers have gained access to tremendous amounts of "omic" data, including for non-model organisms. Phylogenetic information, including species trees, is very limited, unreliable and/or completely unavailable for some poorly studied taxa, thus evidence-based methods are not directly applicable to infer homology. Ebersberger *et al*. (Ebersberger *et al.* 2009) developed the first attempt to circumvent this problem, using a novel hybrid approach (HaMStR) for extraction of homologous sequences from EST/RNA-seq data using a profile Hidden Markov Model (pHMM) (Eddy 1998) based on a similarity search coupled with subsequent RBH derived from re-BLASTing against a reference proteome. The innovative feature of their approach is in the utilization of pHMM as an additional evidence for homology. This architecture incorporates characteristics of multiple sequence alignments (MSA) for user pre-defined core orthologs. Then, a HMM search is performed with each individual pHMM using matching criterion applied to find putative orthologs in the proteome of interest. This method, however, has limitations and weaknesses, such as

i) Proteome training sets composed of phylogeneticaly "meaningful" taxa for construction of core ortholog clusters may not be available,

ii) Identification of informative core ortholog clusters may be somewhat cumbersome due to incomplete and/or low coverage sequencing,

iii) The pHMMs may not contain any relevant compositional or phylogenetic properties about biological sequences that constitute MSA, and

iv) Inability to explicitly identify paralogy limits the use of HaMStR for some evolutionary applications. Hence, homologous clusters inferred from various multiple sequences require further validation to improve confidence in orthology/paralogy classification.

Here, we propose a unique approach to identify false positive homologies detected by heuristic methods, for example HaMStR or InParanoid (Remm *et al.* 2001). Our machine learning method uses phylogenetically-guided inferred homologies to identify non-homologous (false positive) clusters of sequences. This improves the accuracy of heuristic searches, like those that rely on BLAST.

Methods

Library preparation and RNA-seq

For the experimental data set (OD_S) we used 18 Odonata (dragonflies and damselflies) and 2 Ephemeroptera (mayflies) species. Total RNA was extracted from the eye tissues of each taxon using NucleoSpin RNA II columns (Clontech) and reverse-transcribed into cDNA libraries using the Illumina TruSeq RNA v2 sample preparation kit that both generates and amplifies full-length cDNAs. Prepped Ephemeroptera mRNA libraries were sequenced on an Illumina HiSeq 2000 producing 101 bp paired-end reads by the Microarray and Genomic Analysis Core Facility at the Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT, USA, while all Odonata preps were sequenced on a GAIIx producing 72 bp paired-end reads by the DNA

sequencing center at Brigham Young University, Provo, UT, USA. The expected insert sizes were 150 bp and 280 bp respectively. Raw RNA-seq reads were deposited in the National Center for Biotechnology Information (NCBI), Sequence Read Archive, see Additional file 1.

Read trimming and *de novo* transcriptome assembly

The read libraries were trimmed using the Mott algorithm implemented in PoPoolation (Kofler *et al.* 2011) with default parameters (minimum read length = 40, quality threshold = 20). For the assembly of the transcriptome contigs we used Trinity (Grabherr *et al.* 2011), currently the most accurate *de novo* assembler for RNA-seq data (Zhao *et al.* 2011), under the default parameters.

Downstream transcriptome processing

In order to identify putative protein sequences within the Trinity assemblies we used TransDecoder (http://transdecoder.sourceforge.net), the utility integrated into the comprehensive Trinotate pipeline (http://trinotate.sourceforge.net) that is specifically developed for automatic functional annotation of transcriptomes (Haas *et al.* 2013). TransDecoder identifies the longest open reading frames (ORFs) within each assembled DNA contig, the subset of the longest ORFs is then used to empirically estimate parameters for a Markov model based on hexamer distribution. The reference null distribution that represents non-coding sequences is constructed by randomizing the composition of these longest contigs. During the next decision step, each longest determined ORF and its 5 other alternative reading frames are tested using the trained Markov model. If the log-likelihood coding/noncoding ratio is positive and is the highest, this putative ORF with the correct reading frame is retained in the protein collection (proteome). For

more details about the RNA-seq libraries, assemblies and predicted proteomes see Additional file 1.

Construction of *Drosophila* data set

Ten high quality *Drosophila* raw RNA-seq data sets (DROSO) were obtained from NCBI (Additional file 2). First we trimmed the reads using PoPoolation (Kofler *et al.* 2011) and subsampled the read libraries to the size of the smallest (*Drosophila biarmipes*). Then, two additional data sets corresponding to 50% and 10% of the scaled libraries were constructed by randomly drawing reads from the original full-sized libraries. Finally, *de novo* transcriptome assembly and protein prediction were conducted as outlined above for these three data sets. These data sets were used to test whether homology clusters derived from low-coverage RNA-seq libraries contain more false positives.

Gene homology inference

To predict probable homology relationships between proteomes we used the heuristic predictor InParanoid/MultiParanoid based on the RBH concept (Alexeyenko *et al.* 2006; Remm *et al.* 2001). Among various heuristic-based methods for sequence homology detection, OrthoMCL (Li *et al.* 2003) and InParanoid (Remm *et al.* 2001) have been shown to exhibit comparable high specificity and sensitivity scores estimated by Latent Class Analysis (Chen *et al.* 2007), so in the present study we exploited InParanoid/MultiParanoid v. 4.1 for the purpose of simplicity in computational implementation. InParanoid initially performs bidirectional BLAST hits (BBHs) between two proteomes to detect BBHs in the pairwise manner. For this step, we set default parameters with the BLOSUM62 protein substitution matrix and bit score

cutoff of 40 for all-against-all BLAST search. Next, MultiParanoid forms multi-species groups using the notion of a single-linkage. Due to inefficient MultiParanoid clustering algorithm, we had to perform a transitive closure to compile homology clusters for all species together. Transitive closure is an operation performed on a set of related values. Formally, a set S is transitive if the following condition is true: for all values A, B, and C in S, if A is related to B and B is related to C, then A is related to C. Transitive closure takes a set (transitive or non-transitive) and creates all transitive relationships, if they do not already exist. When a set is already transitive, its transitive closure is identical to itself. In the case of the pairwise relationships produced by InParanoid, we constructed orthologous clusters using the notion of transitive closure, where gene identifiers were the values, and homology was the relationship.

For example, our OD_S data set consisted of N = 20 proteomes, so we had to perform $N \times (N - 1)/2 = 190$ pairwise InParanoid queries. A simple transitive closure yielded total 13,998 homology clusters for OD_S. The DROSO data set yielded 20,676, 18,584 and 17,067 homology clusters for 100%, 50% and 10% respectively. Then putative homologous genes were aligned to form individual MSA homology clusters for the subsequent analyses using MAFFT v. 6.864b (Katoh *et al.* 2002) with the "-auto" flag that enabled detection of the best alignment strategy between accuracy- and speed-oriented methods.

Additionally, we utilized HaMStR v. 13.2.3 (Ebersberger *et al.* 2009) under default parameters to delineate putative orthologous sequences in the OD_S proteome sets. 5,332 core 1-to-1ortholog clusters of 5 arthropod species (*Ixodes scapularis, Daphnia pulex, Rhodnius prolixus, Apis mellifera* and *Heliconius melpomene*) for training pHMM were retrieved from the latest version of OrthoDB (Waterhouse *et al.* 2013). We used *Rhodnius prolixus* (triatomid bug) as the reference core proteome because this is the closest phylogenetically related species and

publically available proteome to the Ephemeroptera/Odonata lineage (Meusemann *et al.* 2010). As previously described, each core ortholog cluster was aligned to create MSA using MAFFT and converted into HMM profile using HMMER v. 3.0 (Eddy 2011). BBHs against the reference proteome were derived using reciprocal BLAST.

Construction of ground-truth training sets

The OrthoDB database is one of the most comprehensive collections of putative orthologous relationships predicted from proteomes across a vast taxonomic range (Waterhouse *et al.* 2013). This data is particularly useful for construction of training sets since OrthoDB clusters were detected using a phylogeny-informed approach collated with available functional annotations. Hence, training sets constructed from OrthoDB clusters have the inherent benefit of both an evolutionary and physiological assessment resulting in more precise filtering for false positive homology.

The key to our method was the development of labeled training sets that were used to train supervised machine learning classifiers. Previously, homology clusters were known and annotated in OrthoDB. There were, however, no annotated clusters that represented non-homology clusters from random alignments. Thus, we created and annotated our own set of non-homology clusters through a generative process. We created these clusters in two different manners: randomly aligned sequences and evolving sequences from the homology clusters.

We extracted 5,332 homology (H) clusters from the predefined OrthoDB profile called "single copy in > 70% of species" across the entire arthropod phylogeny in the database, and then aligned them. Non-homology (NH) clusters were generated using: i) the alignment of randomly drawn sequences from the totality of the protein sequences with cluster size sampled

from Poisson(λ), where λ = 44.3056 was estimated as the average cluster size of Hs and ii) by evolving the sequences taken from H clusters. This process of evolving sequences was accomplished by using PAML (Yang 2007) to generate random binary trees for each sequence within a cluster. The discretized number of terminal branches for each random tree was sampled from a normal distribution with mean 50 and a standard deviation of 15. Within each of the clusters, individual sequences were evolved using their respective randomly generated tree using Seq-Gen (Rambaut & Grassly 1997). We used WAG+I (Whelan & Goldman 2001) as the substitution model for the amino acid sequences during the evolving process specifying the number of invariable sites (-i) at 0%, 25% and 50%. Then, to form NH clusters, a single evolved sequence from the terminal branches was selected randomly from each tree. By doing so, we simulated more realistic clusters in which the evolved sequences were diverged enough to be considered as non-homologous to each other.

From the H and NH clusters, two different sets of training, validation and testing partitions were formed. The first set (EQUAL) had an equal number of homology, randomly aligned, 0% invariable-site evolved, 25% invariable-site evolved and 50% invariable-site evolved clusters within the combination of training, validation and testing data sets. The second set (PROP) consisted of 50% of the training set as homology clusters while the remaining half of the training set was composed of equal parts randomly aligned, 0% invariable-site evolved, 25% invariable-site evolved and 50% invariable-site evolved clusters. The combined data sets were then partitioned into training, validation and testing. This was done by randomly sampling from the pool of clusters and assigning 80% of the clusters (8,800) to training, 10% (1,100) to validation and the last 10% (1,100) to testing.

Attribute selection

Ten different attribute features were selected (Table 1) and calculated for individual MSA of putative homology clusters and for training Hs and NHs as well. To identify randomly aligned positions in MSAs, we utilized ALISCORE (Misof & Misof 2009), software based on the principle of parametric Monte Carlo resampling within a sliding window. This approach is more objective and exhibits less conservative behavior contrasted to commonly used non-parametric approaches implemented in GBLOCKS (Castresana 2000; Kuck *et al.* 2010). We expected the number of randomly aligned positions for false positive homologies to be higher than for true homologs. Additionally, several other simple metrics (the number of sequences forming MSAs, alignment length, total number of gaps, total number of amino acid residues and range defined as the difference between longest and smallest sequences within MSAs) were also derived. Overall, incorporation of these attributes into a training set was used to increase the robustness of the performance of the machine learning algorithm. We also obtained amino acid composition for each sequence from each cluster and binned it into four classes according to physicochemical properties of amino acids (charged, uncharged, hydrophobic and special cases), then compositional dispersion was calculated using an unbiased variance estimator corrected for sequence length. Here we assumed that amino acid composition between closely related sequences would be preserved by analogous weak genome-wide evolutionary constraints (Kreil & Ouzounis 2001; Wang & Lercher 2010) and thus have diminished variance.

Machine learning

For detection of false positive homology we utilized different supervised machine learning algorithms in order to learn from the labeled data instances. Supervised machine

learning algorithms take in labeled instances of a particular event as input. From these labeled instances, the algorithm can then learn from the features associated with the instance to perform classification on other, unlabeled instances. A number of different algorithms were used in order to find a model that performed well. Waikato Environment for Knowledge Analysis (WEKA) software (Hall *et al.* 2009) was utilized for training different supervised machine learning classifiers and for evaluating the test data sets. A set of models was trained and compared using the arthropod data set (see **Training data sets** for additional information).

A number of different machine learning algorithms were evaluated. These algorithms included: neural networks, support vector machines (SVMs), random forest, Naive Bayes, logistic regression, and two meta-classifiers. A total of seven models were trained for the arthropod data set. A meta-classifier uses a combination of machine learning algorithms in tandem to perform classification. The two different meta-classifiers utilized stacking with a neural network as the meta-classifying algorithm. Stacking takes the output classifications for all other machine learning algorithms as input and then feeds them into another machine learning algorithm. The learning algorithm that is stacked on the others is then trained and learns which machine learning algorithms it should give more credence when performing classification. One of the meta-classifiers incorporated all the previously mentioned learning algorithms (neural network, SVM, random forest, Naive Bayes, and logistic regression). The other meta-classifier used all the previously mentioned learning algorithms except for logistic regression. All parameters for each machine learning algorithm are summarized in Table 3.

Training

The training data set was used as input to the machine learning model for parameter

selection. For the arthropod data set, 80% of the data were used for training, while 10% of the data was reserved for validation and the last 10% for testing. Machine learning algorithms were utilized to learn from the combination of the H and NH clusters in the data set to differentiate the two. A trained model could then be used to classify unlabeled instances as homologous and non-homologous. There were a total of 8,800 instances in the OrthoDB arthropod data set that were used as a training set for both the PROP and the EQUAL data sets. In the PROP data set, there were 4,378 H and 4,422 NH clusters. In the EQUAL data set, there were 1,753 H and 7,047 NH clusters.

Validation

The validation data sets were used after the model had been trained on the training data set. By using the trained model on the validation set, the efficacy of the model could be seen. 10% of the arthropod data set formed the arthropod validation set. The models trained using the arthropod training set were validated only with the arthropod instances. If the model did not perform adequately on the validation set, different parameters for the machine learning algorithms were modified in an attempt to improve the performance of the models. The re-trained models would then revalidate on their same, respective validation sets. The process was repeated until adequate performance of the learning algorithm was reached. The OrthoDB arthropod validation set consisted of 1,100 instances for both the PROP and EQUAL data sets. The PROP data set had 566 H and 534 NH clusters. The EQUAL data set had 238 H and 862 NH clusters.

Testing

All general steps of our pipeline are summarized in Figure 1 using the example of OD_S processing. Testing data sets were used only after all the models were finished being trained and validated. This is to ensure an honest measure of the predictive capacity of the models because the testing data were never used in order to evaluate how our model was built and to modify the models. The last 10% of the arthropod data set was used as the arthropod test set. The arthropod test set from the OrthoDB contained 1,100 instances for both the PROP and Equal data sets. The PROP data set had 555 H and 545 NH clusters. The EQUAL data set had 207 H and 893 NH clusters.

Performance evaluation

We tested our filtering process by applying the arthropod classifiers trained on the ground-truth data set to the DROSO and OD_S data sets. Unlike the testing sets mentioned in the previous section, the ground-truth for these data sets was unknown. We examined the number of clusters filtered and conducted a manual inspection of a subset of the filtered clusters to verify the removal of only false positive homology clusters. Because there are, to the authors' knowledge, no other post-processing methods for cluster filtering that exist our approach is novel. The filtering processes that do exist are heuristic-based approaches, such as an e-value cutoff, that are built-in modules of the clustering software. Therefore, for comparison, we only examined the number of clusters filtered from the output of InParanoid and HaMStR.

Results and discussion

As can be seen in Table 2 for both the PROP and EQUAL data sets, the arthropod models

all (with the exception of Naive Bayes and SVM) had classification accuracies higher than 96% on the validation set. On the test set, all models (with the exception of Naive Bayes and SVM) had classification accuracies higher than 95%. The algorithms that performed the strongest were the meta-classifiers. The meta-classifier using logistic regression performed best in both the PROP and EQUAL data sets. Comparing the two different data sets, the models perform similarly whether given the PROP or EQUAL data sets. The only exception to this is the Naive Bayes classifier that performs much better (~8% accuracy increase) when given the PROP data set. However, the models trained with the EQUAL data sets were slightly better in terms of accuracy (Figure 2). In the arthropod models, we varied the size of the training set (from 1% to 100% of training instances). The validation set accuracy of the meta-classifier with logistic regression plateaued and slowed growth after training on 5% or more of the training instances. Before this, their classification accuracies of all models were erratic with both increases and decreases as the training set size increased. The models behave differently when given varied amounts of data to train on (Table 2). All models except for Naive Bayes increased in accuracy as the training data grew. Logistic regression and the meta-classifier with logistic regression required the least amount of training data before they started to plateau. Additionally we tested which features were the most meaningful for classification using meta-classifier with logistic regression (Fig. 3). The "number of gaps" feature provided the best accuracy when 100% of instances were used. Since increased indel events are accumulated over longer evolutionary time periods, the inferred MSAs from such highly diverged sequences with lost signatures of common ancestry are expected to have multiple gaps. Moreover, clusters prone to large amounts of missing data will be classified as NH using this feature. Similar accuracy levels were achieved for the four amino acid composition and number of amino acids features. As we mentioned

earlier, selection forces may preserve amino acid composition especially through the action of purifying selection (Hughes 2010) making these features useful for H vs. NH cluster discrimination. Other features, except for Aliscore that exhibited an intermediate accuracy, had accuracy < 80%, which might be explained by the fact that these features are less biologically meaningful.

Lower coverage data sets are often used when performing transcriptomic and evolutionary analyses especially on non-model organisms. For instance, in a recent paper (Misof *et al.* 2014) the authors inferred a phylogeny of many insect species using relatively small RNA-seq library sizes averaging at ~ 3Gb (Additional file 2) compared to *Drosophila* data sets (Additional file 3). We expected the number of false positive clusters to increase with the decreasing sequencing depth. In order to examine this, three DROSO data sets were tested for the presence of false positives using the meta-classifier with logistic regression trained on the EQUAL arthropod data set. Indeed, we found that the number of false positive homology clusters increased in the subsampled DROSO data sets (15.7%, 17.8% and 29.9% for 100%, 50% and 10% DROSO data sets respectively). These subsampled data sets allowed us to see the results that are common when homology clustering is performed on small libraries. Applying the filtering process to the InParanoid and HaMStR OD_S clusters resulted in many removed clusters (Table 4), implying that heuristic-based methods have increased rates of false positives. For filtering, we only used the meta-classifier with logistic regression. The removal of many clusters showed the overall poor quality of many of the putative homology clusters (for comparison between homology and false-positive homology clusters see Figure 4). This was expected due to the low quality transcriptome assembly that was caused by sequencing depth in addition to biological factors such as interspecific differential expression. The filtering process

preserved higher quality clusters and finished almost instantly resulting in huge time savings when compared to manually curating the clusters. Overall our method can be applied to filter homology clusters derived from closely related (e.g. *Drosophila* species) as well as highly diverged taxa (e.g. Odonata species). We also note that the trimming procedure behaves more conservatively with increasingly diverged sequences.

Conclusions

We have demonstrated a machine learning method that can be used to differentiate homology and non-homology clusters based on characteristics of known good and bad clusters. These results can be seen in our trained models' ability to achieve high classification accuracy on the test data sets as well as by examining the number of clusters that were removed from the experimental OD_S data set. We developed a training set of known good and bad clusters that was previously unavailable and made supervised machine learning impossible. Using a feature set that we developed, we tested various machine learning algorithms and found that when trained on our training data sets that the meta-classifier with logistic regression consistently outperformed all other models and performed just as well as the meta-classifier without logistic regression.

Applications of our method were also seen as we applied them to other data sets. Our method was especially useful when applied to the OD_S data set, by filtering out many clusters with false positive homology. We showed that our method is effective in settings where non-model organisms are being studied and the transcriptome assembly quality is low primarily due to low coverage sequencing or partial RNA degradation.

This paper has demonstrated the usefulness of machine learning in finding homology

clusters by quickly removing low quality clusters without using any additional heuristics. The clusters that are retained can then be used later in higher quality phylogeny reconstruction and/or other analyses of gene evolution. In the future, we aim to explore machine learning approaches to clustering sequences more deeply to produce more refined and reliable homology clusters.

Authors' contributions

AS proposed the idea and designed the experiments. MSF trained the machine learning algorithms. MSF and NOJ analyzed the data. MJC and SMB provided reagents/data/analysis strategies. AS, MSF and NOJ wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

We thank Gavin J. Martin and Nathan P. Lord for the generation of sequence data, T. Heath Ogden for providing specimens and Eric Ringger for his help with machine learning model selection and valuable discussion. We also thank the National Science Foundation for funding this research in the form of a grant awarded to both SMB and MJC (IOS-1265714).

References

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

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

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* **2**, e383.

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361-375.

Dessimoz C, Gabaldon T, Roos DS*, et al.* (2012) Toward community standards in the quest for orthologs. *Bioinformatics* **28**, 900-904.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**, 157.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755-763.

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195.

Gabaldon T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14**, 360-366.

Grabherr MG, Haas BJ, Yassour M*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652.

Haas BJ, Papanicolaou A, Yassour M*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512.

Hall M, Frank E, Holmes G*, et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10-18.

Hughes AL (2010) Evolutionary conservation of amino acid composition in paralogous insect vitellogenins. *Gene* **467**, 35-40.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066.

Kofler R, Orozco-terWengel P, De Maio N*, et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**, e15925.

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-338.

Kreil DP, Ouzounis CA (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* **29**, 1608-1615.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. *Brief Bioinform* **12**, 379-391.

Kuck P, Meusemann K, Dambach J*, et al.* (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool* **7**, 10.

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189.

Meusemann K, von Reumont BM, Simon S*, et al.* (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* **27**, 2451-2464.

Misof B, Liu S, Meusemann K*, et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767.

Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* **58**, 21-34.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901.

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13**, 235-238.

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052.

Wang GZ, Lercher MJ (2010) Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evol Biol* **10**, 263.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41**, D358-365.

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

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591.

Zhao QY, Wang Y, Kong YM*, et al.* (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12 Suppl 14**, S2.

Figures and Tables

Figures

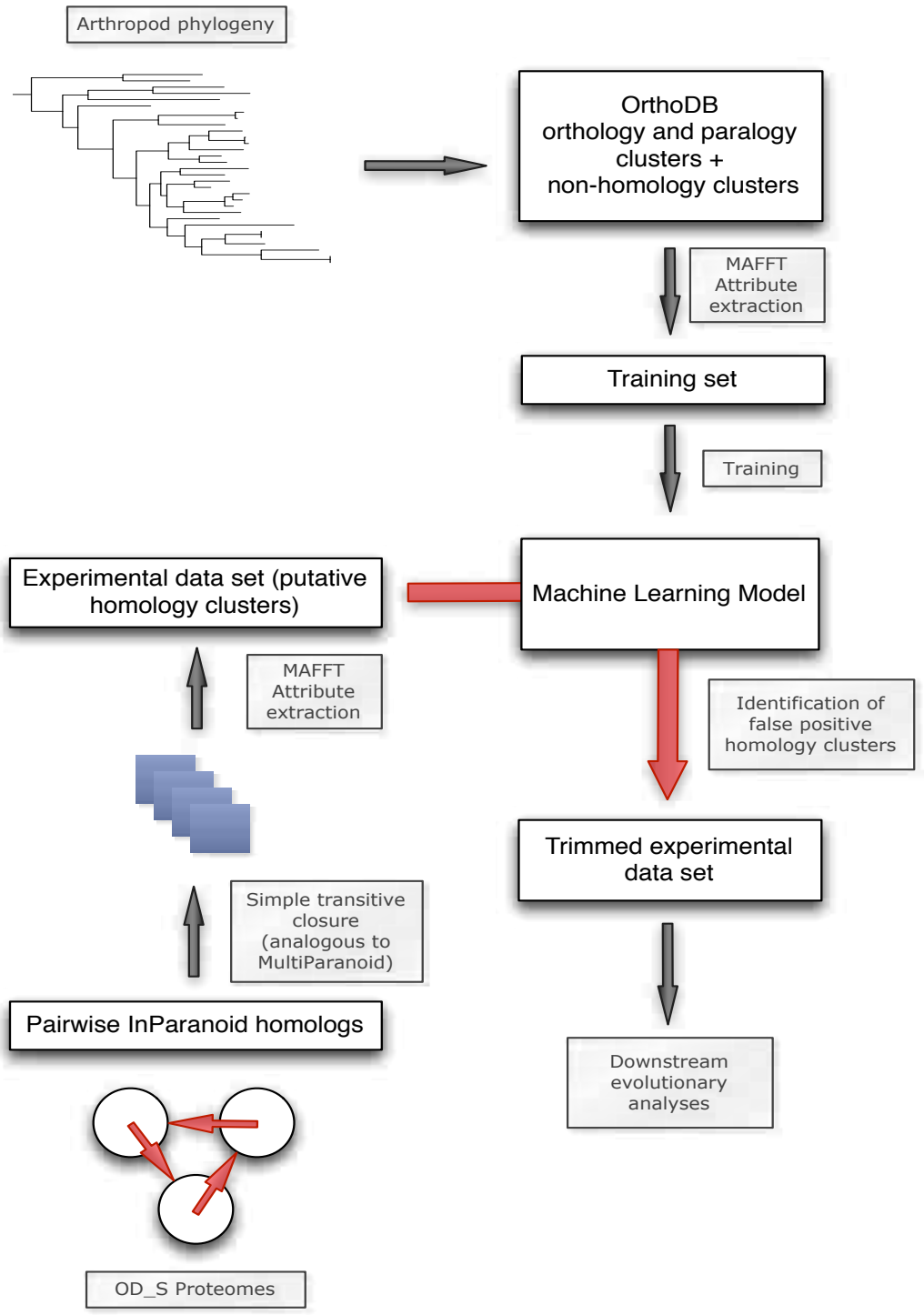


Figure 1. A diagram of the workflow. This figure shows the different steps that were used in

developing our machine learning model. Arthropod phylogeny was generated in previous studies and deposited in OrthoDB. These sequences were then gathered from OrthoDB and used as our orthology and paralogy clusters. They were combined with generated non-homology clusters. The combination represents our training data set used to train the machine learning algorithms. The experimental data were assembled with proteins inferred from the assemblies. InParanoid was then used to identify putative homologs. Once putative homologs were identified they were input into the trained machine learning algorithms for classification and subsequent cluster trimming.

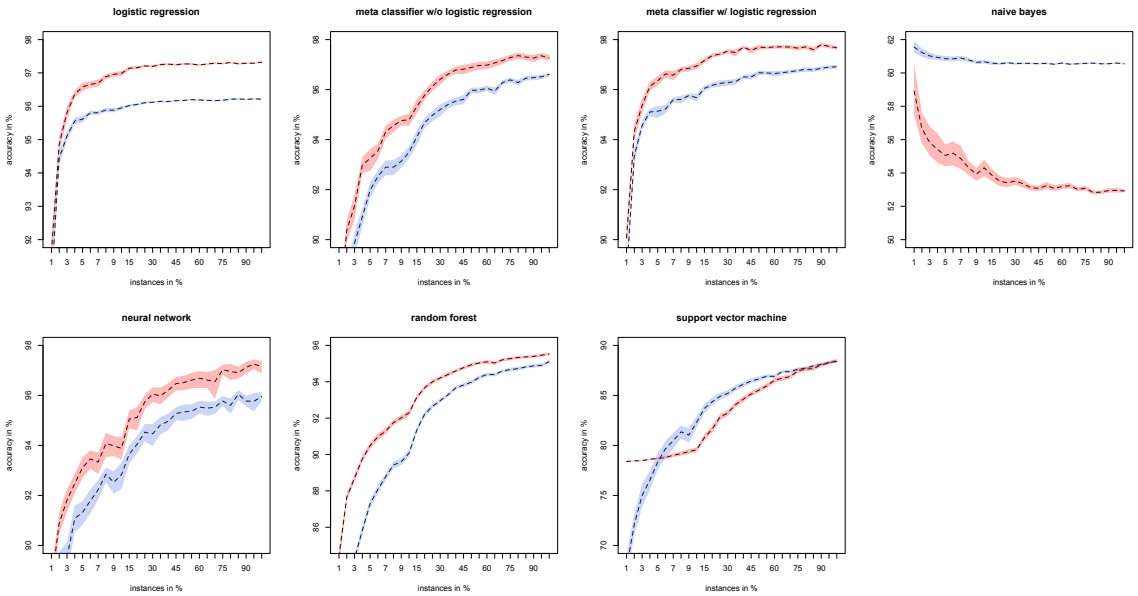


Figure 2. Bootstrapping results for the machine learning models. Bootstrapping was conducted using 100 replicates for each classifier. Error envelopes can also be seen for each classifier. Except for Naive Bayes, as the percentage of total training instances used during learning increases accuracy increases and the error envelope decreases.

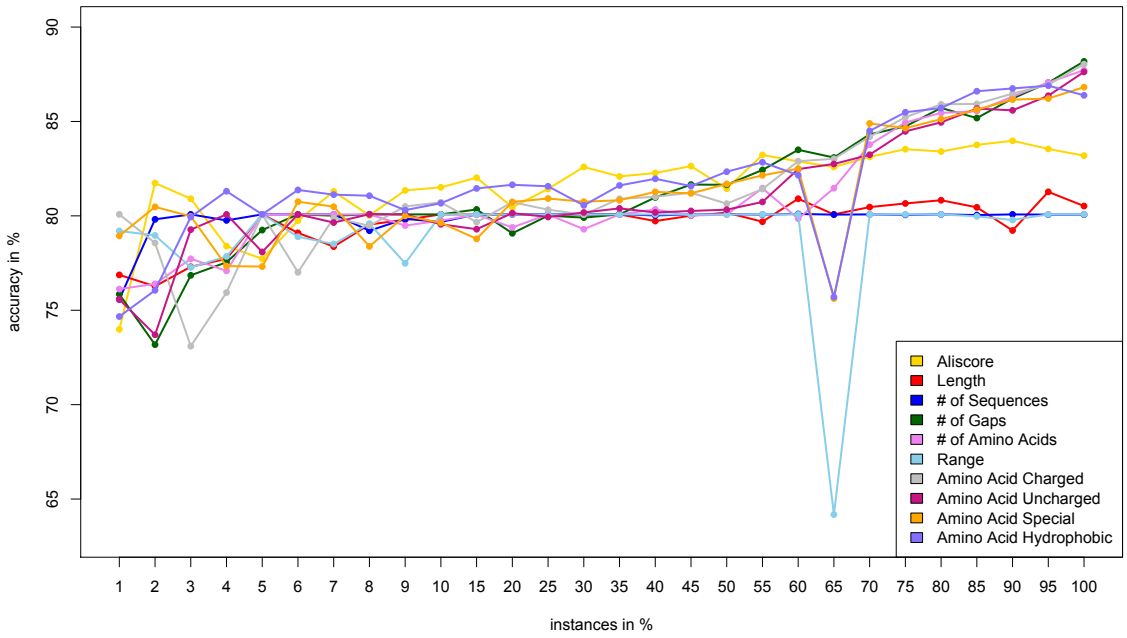


Figure 3. Accuracy curves for individual features (EQUAL training data set) using meta-classifier w/ logistic regression. The number of gaps, amino acid composition and number of amino acids features exhibit better predictive accuracy.

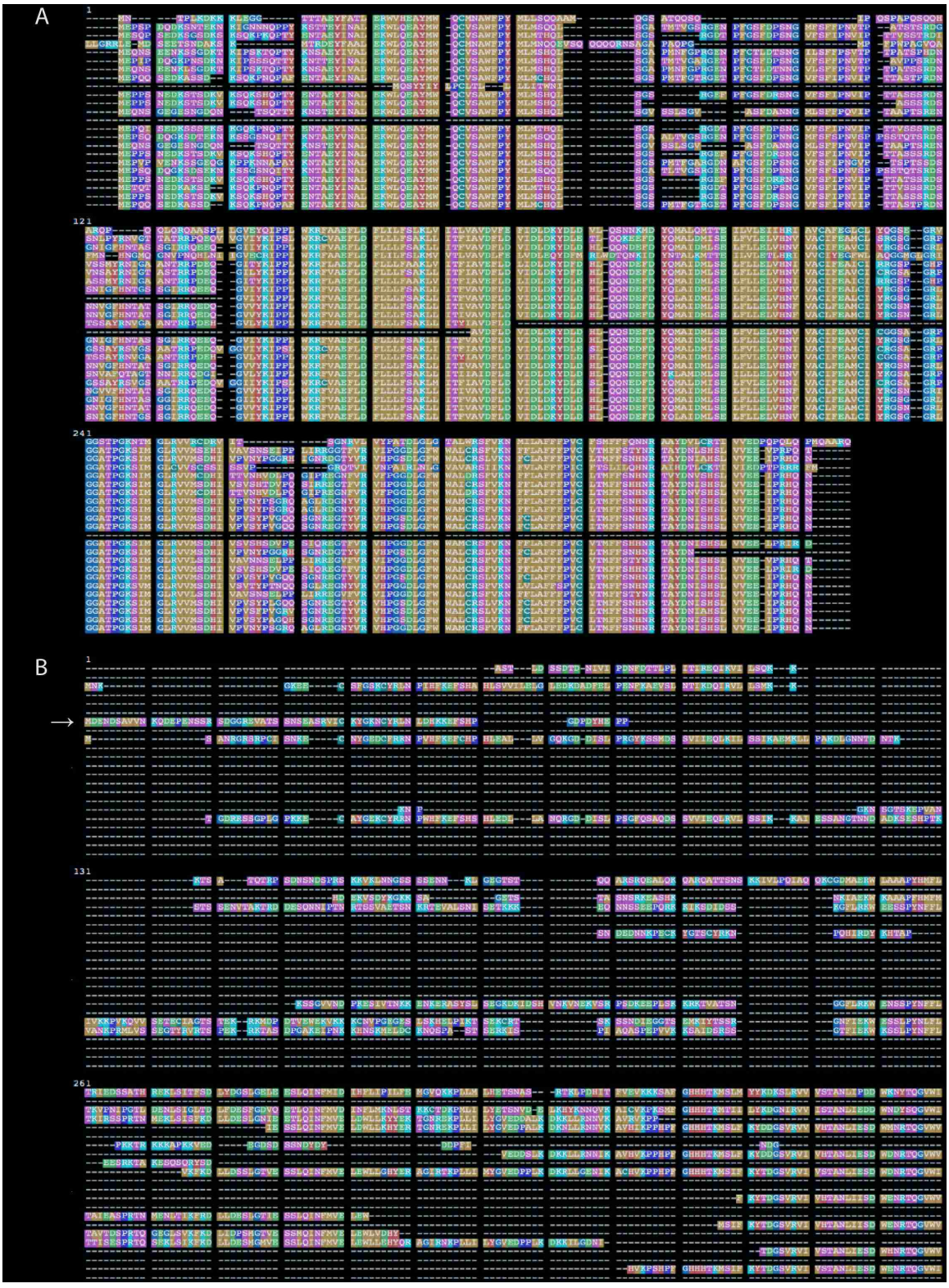


Figure 4. Examples of a high quality homology (A) and false-positive homology (B) clusters (OD_S data set) classified by meta-classifier w/ logistic regression. All sequences within the

homology cluster (A) belong to one protein family (FAM81A1-like protein). The sequence in the false-positive homology cluster indicated by the arrow represents Aprataxin and PNK-like factor whereas other sequences represent tyrosyl-DNA phosphodiesterase.

Tables

Table 1. All Features that were used in order to train the machine learning algorithm. Each of these features was calculated for each of the clusters.

Feature	Description
Aliscore	The number of positions identified by Aliscore as randomly aligned
Length	The length of the alignment
# of Sequences	The number of sequences in the alignment
# of Gaps	Number of base positions marked with a gap
# of Amino Acids	Number of amino acids in the alignment
Range	Longest non-aligned sequence length minus shortest non-aligned sequence length
Amino Acid Charged	Standard deviation for the proportions of amino acids in the charged class for each sequence
Amino Acid Uncharged	Standard deviation for the proportions of amino acids in the uncharged class for each sequence
Amino Acid Special	Standard deviation for the proportions of amino acids in the non-charged and non-hydrophobic class for each sequence
Amino Acid Hydrophobic	Standard deviation for the proportions of amino acids in the hydrophobic class for each sequence

Table 2. Summary of arthropod machine learning model performance. This table shows the performance of each of the different learning algorithms that were trained, validated, and tested with the OrthoDB arthropod gene clusters.

	OrthoDB Arthropod EQUAL		**OrthoDB Arthropod PROP**	
Algorithm	**Validation**	**Testing**	**Validation**	**Testing**
Neural Network	97.1815%	96.8153%	97.5452%	96.5423%
Suppor Vector Machine (SVM)	89.1351%	88.0801%	88.0668%	88.2621%
Random Forest	98.1362%	95.9054%	97.8748%	95.5414%
Naive Bayes	53.0628%	52.5023%	61.2229%	60.3276%
Logistic Regression	96.5905%	97.2702%	96.3064%	96.3603%
Meta-Classifier w/o Logistic Regression	98.5112%	98.3621%	98.5907%	96.8153%
Meta-Classifier w/ Logistic Regression	98.6362%	97.7252%	98.5680%	97.5432%

Table 3. The machine learning parameters used for each of the different algorithms in WEKA.

Algorithm	Parameters
Neural Network	weka.classifiers.functions.MultilayerPerceptron -L 0.1 -M 0.05 -N 3000 -V 0 -S 0 -E 40 -H a
Support Vector Machine (SVM)	weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
Random Forest	weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1
Naive Bayes	weka.classifiers.bayes.NaiveBayes
Logistic Regression	weka.classifiers.functions.Logistic -R 1.0E-8 -M -1
Meta-Classifier w/o Logistic Regression	weka.classifiers.meta.Stacking -X 10 -M "weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a" -S 1 -B "weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1" -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0""
Meta-Classifier w/Logistic Regression	weka.classifiers.meta.Stacking -X 10 -M "weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a" -S 1 -B "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1" -B "weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a" -B "weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1" -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0""

Table 4. Summary of InParanoid and HaMStR cluster filtering. The number of clusters that were kept and removed for the OD_S clusters from InParanoid and HaMStR. Filtering was accomplished using the meta-classifier w/ logistic regression model trained on the EQUAL data set.

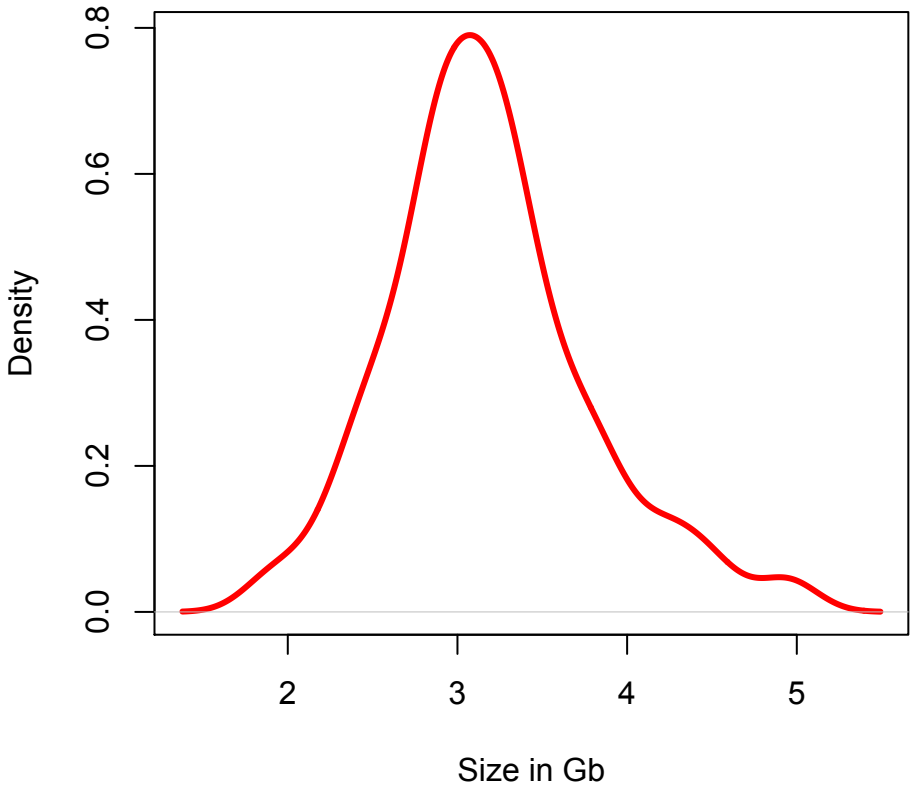
		Kept	Removed
Odonata	**InParanoid**	10500	3497
	HaMStR	1231	896

Additional files.

Additional file 1. Summary of OD_S RNA-seq libraries

Library	Reads Before Trimming	Reads After Trimming	N50	Max Contig Length	Min Contig Length	N of Contigs	N of Peptides (TransDecoder)	SRA ID (NCBI)
OD07_Cordulegaster_maculata	7207518	6819629	983	16508	201	28163	11877	SRR2164542
OD08_Anax_junius	6267456	5978119	807	9457	201	19987	8519	SRR2164543
OD10_Hetaerina_americana	6447244	6057989	1141	10815	201	34384	13373	SRR2164551
OD11_Ischnura_verticalis	6183018	5790475	879	7894	201	27001	10568	SRR2164552
OD12_Gomphus_spicatus	6498099	6273168	1502	11612	201	37936	10611	SRR2157378
OD13_Nehalennia_gracilis	5894197	5516510	1027	11774	201	33766	12694	SRR2157379
OD18_Chromagrion_conditum	7607629	7189745	1188	8671	201	30453	9256	SRR2157380
OD25_Stylurus_spiniceps	6840281	6597769	1249	23861	201	37436	13568	SRR2157381
OD28_Neurocordulia_yamaskanensis	6410925	6061801	1261	15978	201	34984	12905	SRR2157382
OD36_Argia_fumipennis_violacea	5955971	5600701	1076	11410	201	35049	12754	SRR2157383
OD42_Archilestes_grandis	8736454	8367974	1179	9315	201	34318	12285	SRR2164544
OD43_Hetaerina_americana_2	3585483	3411596	738	7343	201	24192	7318	SRR2164545
OD44_Enallagma_sp	5646110	5370720	940	6446	201	27135	8085	SRR2157367
OD45_Libellula_forensis	5591547	5383248	1125	13425	201	31962	11352	SRR2164546
OD46_Libellula_saturnata	5961628	5717528	1397	13986	201	35045	13326	SRR2164547
OD62_Ischnura_hastata	10080263	9551907	1777	11200	201	40154	13651	SRR2164548
OD64_Anax_junius_2	9657180	9195840	1133	21566	201	30833	13117	SRR2157371
OD_Ischnura_cervula	7105927	6702900	1156	15621	201	40741	14253	SRR2157372
R_E001_Baetis_sp	16352942	16113853	1786	10772	201	30517	16743	SRR2164549
R_E006_Epeorus_sp	13846765	13701079	1303	23453	201	45886	16782	SRR2164550

Additional file 2. Density estimation of RNA-seq base coverage used in (Misof *et al.* 2014)



Additional file 3. Summary of DROSO RNA-seq libraries

Drosophila Species	NCBI ID	# of bases (in Gb)	Platform
D. ananassae	SRR166825	13.6	Illumina HiSeq 200 PE
D. biarmipes	SRR346718	6.4	Illumina HiSeq 200 PE
D. ficusphila	SRR346748 SRR346751	12.1	Illumina HiSeq 200 PE
D. mauritiana	SRR1560444	7.7	Illumina HiSeq 200 PE
D. melanogaster	SRR1197414	9.6	Illumina HiSeq 200 PE
D. miranda	SRR899848	13	Illumina HiSeq 200 PE
D. mojavensis	SRR166833	11.1	Illumina HiSeq 200 PE
D. pseudoobscura	SRR166829	15.1	Illumina HiSeq 200 PE
D. simulans	SRR166816	17.2	Illumina HiSeq 200 PE
D. virilis	SRR166837	15.1	Illumina HiSeq 200 PE
D. yakuba	SRR166821	13	Illumina HiSeq 200 PE