A Phylogenomic Approach to Resolve the Basal Pterygote Divergence

Sabrina Simon,* Sascha Strauss,† Arndt von Haeseler,† and Heike Hadrys†‡

*Institute of Ecology & Evolution, Stiftung Tierarztliche Hochschule Hannover, Hannover, Germany; †Center for Integrative Bioinformatics Vienna, Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, University of Veterinary Medicine, Vienna, Austria; and ‡Department of Ecology and Evolutionary Biology, Yale University

One of the most fascinating Bauplan transitions in the animal kingdom was the invention of insect wings, a change that also contributed to the success and enormous diversity of this animal group. However, the origin of insect flight and the relationships of basal winged insect orders are still controversial. Three hypotheses have been proposed to explain the phylogeny of winged insects: 1) the traditional Palaeoptera hypothesis (Ephemeroptera + Odonata, Neoptera), 2) the Metapterygota hypothesis (Ephemeroptera, Odonata + Neoptera), and 3) the Chiastomyaria hypothesis (Odonata, Ephemeroptera + Neoptera). Neither phylogenetic analyses of single genes nor even multiple marker systems (e.g., molecular markers + morphological characters) have yet been able to conclusively resolve basal pterygote divergences. A possible explanation for the lack of resolution is that the divergences took place in the mid-Devonian within a short period of time and attempts to solve this problem have been confounded by the major challenge of finding molecular markers to accurately track these short ancient internodes. Although phylogenomic data are available for Neoptera and some wingless (apterygote) orders, they are lacking for the crucial Odonata and Ephemeroptera orders. We adopt a multigene approach including data from two new expressed sequence tag projects—from the orders Ephemeroptera (Baetis sp.) and Odonata (Ischnura elegans)—to evaluate the potential of phylogenomic analyses in clarifying this unresolved issue. We analyzed two data sets that differed in represented taxa, genes, and overall sequence lengths: maxsp (15 taxa, 125 genes, and 31,643 amino acid positions) and maxgen (8 taxa, 150 genes, and 42,541 amino acid positions). Maximum likelihood and Bayesian inference analyses both place the Odonata at the base of the winged insects. Furthermore, statistical hypotheses testing rejected both the Palaeoptera and the Metapterygota hypotheses. The comprehensive molecular data set developed here provides conclusive support for odonates as the most basal winged insect order (Chiastomyaria hypothesis). Data quality assessment indicates that proteins involved in cellular processes and signaling harbor the most informative phylogenetic signal.

Introduction

Insects are the most diverse animal group on earth and dominate every ecosystem except the benthic zone (Grimaldi and Engel 2005). The winged insects—Pterygota—account for more than 98% of the class Insecta (Grimaldi and Engel 2005). According to fossil records, flying insects originated in the Early Carboniferous period (ca. 320 Ma), whereas a DNA-based study suggested an origin in the mid-Devonian (ca. 387 Ma; Gaunt and Miles 2002). A recent analysis of Engel and Grimaldi (2004) suggested that the origin of insect wings occurred coincident with the development of arborescence and agreed with the molecular estimates of Gaunt and Miles (2002). With the invention of the wings, insects were able to invade every ecosystem, escape predators, and exploit scattered resources, resulting in rapid radiations into vast numbers of species (Hennig 1969). Considering the tremendous impact this change produced, the evolution of the flying insects is one of the most fascinating questions in evolutionary biology. Martynov (1925) was the first to distinguish two groups of winged insects based on wing function—Palaepoptera and Neoptera. He assumed the inability to fold back the wings, as seen in Ephemeroptera and Odonata, to be an ancestral condition and therefore called them Palaepoptera (old wings) in contrast to those with this ability, which he called Neoptera (new wings). The monophyly of Palaepoptera has been controversial ever since. In contrast to the accepted monophyly of Neoptera, the so-called “Palaepoptera Problem” is one of the unsolved mysteries in insect systematics.

Key words: basal pterygote divergence, Palaepoptera, Metapterygota, Chiastomyaria, phylogenomics, expressed sequence tags.

E-mail: sabrina.simon@ecolevol.de.

doi:10.1093/molbev/msp191
Advance Access publication August 27, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
same genes.

blue and the resulting basal pterygote order in red. Below are different molecular studies listed supporting one of the three hypotheses partly using the same genes.

Odonata support the Chiastomyaria theory (Boudreaux 1979). Moreover, the wing base structure of the Odonata and the remaining pterygote orders show significant differences in appearance and function, for example, wing flapping in Odonata is promoted by the direct flight muscles, whereas in Ephemeroptera and Neoptera, it is promoted by indirect flight muscles (Ninomiya and Yoshizawa 2009). The difficulties in establishing homology of the wing base structure between the Odonata and other Pterygota resulted in an extreme interpretation of Matsuda (1970, 1981) and La Greca (1980). They concluded that the wing base structure in odonates is so different that it cannot be homologized with that of Ephemeroptera and Neoptera. However, the monophyly of Pterygota is now well established through both morphology and molecular data (e.g., Kristensen 1991; Wheeler et al. 2001; Grimaldi and Engel 2005; Kjer et al. 2006; von Reumont et al. 2009). Recently, Ninomiya and Yoshizawa (2009) established the homology of the wing base structures between the Odonata, Ephemeroptera, and Neoptera. Based on wing base morphology, they almost unambiguously determined that there is a single origin of insect wings and flight but not able to contribute further on the basal diversification of Pterygota.

Establishing a sound phylogenetic hypothesis for the origin of insect wings based on wing base structure and the wing folding mechanism remains crucial. But why is the so-called Palaeoptera Problem not resolved despite the advances in molecular systematics? Whitfield and Kjer (2008) pointed out that the “ancient rapid radiation” is a major contributing factor in the inability to resolve insect relationships with molecular data. Due to short ancient internodes, connecting the taxonomic groups, inadequate molecular data sets, conflicting results within or among data sets, and an overall weak phylogenetic signal is observed in many pterygote phylogenetic studies (Wheeler et al. 2001; Ogden and Whiting 2003; Kjer et al. 2006; Misof et al. 2007; von Reumont et al. 2009). In addition, one major challenge is to find useful molecular markers to accurately track these short ancient internodes. For the reconstruction of an “accurate” phylogeny, molecular marker systems are required which have kept pace with specification but slow enough to have transferred the phylogenetic signal to the present (Regier and Shultz 1998). Unfortunately, the rationale behind the selection of certain molecular markers is not always clear, and discrepancies and incongruence between individual gene trees may result in unresolved phylogenetic trees (Wheeler et al. 2001; Kjer et al. 2006). Thus, phylogenetic analyses of single genes and even multiple marker systems have not yet conclusively resolved the basal pterygote diversification. It is therefore conceivable that resolution of these relationships may require not only large amounts of sequence data but also an assessment of data quality and quantity. Several studies have shown that analyzing a large number of genes simultaneously helps to infer unresolved issues in deep metazoan relationships (e.g., Philippe, Lartillot, and Brinkmann 2005; Savard et al. 2006; Roeding et al. 2007; Dunn et al. 2008). Moreover, simulations and studies based on real data have shown that trees based on concatenated alignments provide better resolution for a particular topology than consensus gene trees—known as “supertree” approaches (Rokas, Williams, et al. 2003; Gadagkar et al. 2005; Savard et al. 2006). However, there is still a controversy about phylogenetic reconstructions derived from supertree versus “supermatrix” approaches (e.g., Gatesy et al. 2004; Wilkinson et al. 2007). Both methods have demonstrated strengths and weaknesses, and some promising new approaches are addressing the existing problems. For example, for the supertree method, the recent proposal of a maximum likelihood (ML) approach forms an important idea for future phylogenetic inferences from genomic data (Steel and Rodrigo 2008; Cotton and Wilkinson 2009). Also the implementation of new methods, for example, Bayesian estimation of species trees (Edwards et al. 2007; Liu and Pearl 2007), to simultaneously estimate gene trees and species trees from multilocus data using a coalescent framework has been shown to be very efficient in cases of recent speciation (Edwards et al. 2007; Belfiore et al. 2008; Wiens et al. 2008). All these phylogenomic approaches have one problem in common; although the stochastic error is dramatically reduced by using a large number of data, they are not protected against systematic errors (Phillips et al. 2004; Delsuc et al. 2005).
Furthermore, systematic bias can be reinforced by increasing the number of characters resulting in a highly supported but incorrect tree (Felsenstein 1978; Jeffroy et al. 2006). Long-branch attraction (LBA) coupled with taxon sampling, phylogenetic reconstruction methods and base composition bias are all factors that are known to cause systematic errors and to be potential pitfalls when attempting to recover “the true evolutionary history of species” (Zwickl and Hillis 2002; Phillips et al. 2004; Brinkmann et al. 2005; Delsuc et al. 2005; Philippe, Delsuc, et al. 2005).

With the aim of addressing the origin of flying insects, we generated and analyzed expressed sequence tag (EST) data from the two basal orders of winged insects—from a mayfly (Ephemeroptera, Baetis sp.) and a damselfly (Odonata, Ischnura elegans). EST data provide a comprehensive random sample of protein-coding genes and an economic way to produce a large number of sequences for phylogenetic analysis of “nonmodel” species, for which genome sequence projects are not yet available.

Although EST data collection is increasing due to the tremendous recent advances in sequencing techniques and as an optimal source for multigene approaches, ESTs from representatives of the basal winged insect orders are still scarce.

Although ESTs are a promising tool to resolve deep phylogenetic questions, there are still necessary precautions to take when handling EST data sets. The complex nature of genome evolution including gene loss, duplications, expansion of gene families, and functional diversification consequently requires assignment of gene orthology when using ESTs as a source for phylogenetic analyses (Hughes et al. 2006). Furthermore, ESTs represent a snapshot of gene expression within a given set of tissue, developmental stages, and environmental conditions (Rudd 2003), and the overlap of genes in the taxa may be very limited (Hughes et al. 2006).

In this study, we have assigned gene orthology using the new search algorithm hidden Markov model-based search for orthologs using reciprocity (HaMStR; Ebersberger et al. 2009) and constructed two alignments to evaluate the new search algorithm hidden Markov model–based phylogeny (http://www.deep-phylogeny.org/hamstr/) with HaMStR (Ebersberger et al. 2009; http://www.deep-phylogeny.org/hamstr/). For each taxon, identification of orthologs was carried out using the HaMStR approach and only a single clustering step was performed. The data sets were subjected to different statistical and evolutionary analyses to test the hypotheses and to gain more insights into the origin of flying insects. The phylogenetic information contained within the different protein-coding genes and represented within the data sets was also assessed.

Materials and Methods

**cDNA Library Construction, EST Processing, and Sequence Alignment**

Specimens were stored in RNAlater (Qiagen) at −80 °C before RNA extraction. Total RNA of *Baetis* sp. was extracted four times from two larval specimens simultaneously using Qiagen RNEasy kits and pooled afterward. Total RNA of *I. elegans* was extracted from one adult specimen using Qiagen RNEasy kits. The two RNA samples were precipitated with 0.1Vol NaAC in diethylpyrocarbonate and 2.5Vol 100% ethanol for later construction of cDNA libraries.

The Creator SMART cDNA Library Construction Kit (Clontech) and the Trimmer Kit (Evrogen) were used for the construction of the normalized cDNA libraries following the manufacturers’ instructions. Modifications to the protocol were made concerning the cloning vector: pal32 (Evrogen) was used for directional cloning with insertion between two SfiI sites. Plasmids were transferred via electroporation to *Escherichia coli* (strain DH10B, Invitrogen). Plasmids were isolated using the method of Hecht et al. (2006) and 5′ end sequenced using BigDye V3 (ABI) and 3730XL capillary sequencer systems (ABI).

The program Lucy (Chou and Holmes 2001) removed vector contaminations in the raw sequences. Additionally, all sequences were screened for contamination by comparing them to the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) with CrossMatch (http://www.phrap.org/phredphrapconsed.html) and SeqClean (http://compbio.dfci.harvard.edu/cgi/software). The latter program was also used to remove poly(A) tails. Subsequently, ESTs with less than 100 nt were discarded. Repetitive elements were soft masked using RepeatMasker (Smit AF, Hubley R, Green P, unpublished data) and Repbase (Jurka et al. 2005).

ESTs for each species were clustered and assembled using TIGR gene indices clustering tools (Pertea et al. 2003). The resulting EST contigs were quality clipped with Lucy, and again sequences of less than 100 nt were removed. Afterward the quality clipped, sequences were clustered a second time.

*Baetis* sp. ESTs have been deposited in the EMBL Nucleotide Sequence Database with GenBank accession numbers FN198828–FN203024 and *I. elegans* ESTs with GenBank accession numbers FN215340–FN219556.

Individual EST contigs were compared with the NCBI nonredundant protein database using BlastX (Altschul et al. 1997). The protein sequences of the best 25 Blast hits per contig were extracted from the database and aligned to the contig separately using GeneWise (Birney et al. 2004). The description of the protein sequence resulting in the highest GeneWise alignment score was adopted as tentative annotation.

Further ESTs were downloaded from NCBI’s database of “Expressed Sequence Tags.” A processing analog to the procedure explained above was applied except that Lucy was not used for vector screening and quality clipping, and only a single clustering step was performed.

We included 25 pterygote and 3 apterygote specimens in our data set (supplementary table S1, Supplementary Material online). For each taxon, identification of orthologous genes was carried out using the HaMStR approach (Ebersberger et al. 2009; http://www.deep-phylogeny.org/hamstr/) with *Anopheles gambiae*, *Apis melifera*, *Drosophila melanogaster*, *Homo sapiens*, and *Aedes aegypti* as core reference taxa and a reblast of the candidate
EST contigs against A. melifera as a reference proteome. Overall, our core ortholog set encompassed 3,096 clusters of orthologous genes, which were used to assign EST contigs to individual genes.

Because not all the 3,096 genes were present in the EST contigs of each taxon, a concatenation of all gene alignments would have resulted in a substantial amount of missing data. We therefore used a PERL script (kindly provided by Ingo Ebersberger and available upon request; ingo.ebersberger@univie.ac.at) that automatically analyzes the amount of missing data for different combinations of taxa and genes. We have chosen two data sets representing different taxa and genes and a diverse proportion of missing data. We decided to perform all analyses with both data sets to make our results more robust. As a selection criterion for the data sets, we imposed that Baetis sp., I. elegans, and at least one apterygote taxon were present in each set. One data set (named maxspe) comprised 15 species and 125 genes with 18% missing data and a second (named maxgen) comprised 8 species, 150 genes, and 11% missing data.

Sequences were aligned with MAFFT (Katoh et al. 2005) using the options—maxiterate 1,000 and—localpair. Afterward we concatenated the alignments to generate one super alignment per data set (supplementary table S2, Supplementary Material online for list of represented genes).

Phylogenetic Analyses of Concatenated Data

Both alignments (maxspe, maxgen) were checked for putatively randomly similar sections using ALISCORE (Misof B and Misof K 2009). We applied a sliding-window size (v = 6) with the BLOSUM62 matrix and function -e (option for EST data with lots of missing data section). After the exclusion of putative randomly similar sections using PAUP*4.0b (Swofford 2002), we determined the best fitting model of protein sequence evolution with ProtTest 1.4 (Abascal et al. 2005), which was used in subsequent phylogenetic analyses.

Maxspe and maxgen were treated equally in all following steps of phylogenetic and statistical analyses. Tests of the three alternative phylogenetic hypotheses at the base of the Pterygota were accomplished by using the approximately unbiased (AU) test, Kishino–Hasegawa (KH), Shimodaira–Hasegawa (SH), weighted Kishino–Hasegawa (WKH), and weighted Shimodaira–Hasegawa (WHS) tests as implemented in CONSEL (Shimodaira and Hasegawa 2001). First, alternative tree topologies were reconstructed by using Garli 0.96b8 (Zwickl 2006) under default parameters. Heuristic searches were conducted assuming the WAG (Whelan and Goldman 2001) model of amino acid (aa) sequence evolution and a Λ-model of rate heterogeneity (Gu et al. 1995), with four classes of variable sites and one class of invariable sites (4Γ + 1). As the monophyly of the major groups is not disputed, we put a topological constraint according to the three phylogenetic hypotheses on the tree search to identify the highest likelihood topologies that satisfied a given hypothesis. In addition, we constrained the monophyly of Paraneoptera and Holometabola in the maxgen data set (e.g., Hennig 1981; Yoshizawa and Saigusa 2001; Kaestner 2003; Beutel and Pohl 2006). PAUP* was used to produce a file with the site-wise log likelihoods of alternative trees. The resulting files were summarized to a single file that served as input for CONSEL to calculate the P value for each alternative phylogenetic hypothesis.

In addition to the constrained analyses, searches in the absence of topological constraints were carried out. For this purpose, ML analyses were performed with the Phreads-parallelized version of RAxML 7.0.4 (Stamatakis 2006) under a rapid bootstrap analysis (-f a) and the PROTMIX-WAG model. The branching support was assessed by 1,000 bootstrap replicates. Bayesian inference (BI) analyses were performed using a compiled parallel version of MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Altekar et al. 2004) with two parallel runs under the WAG + 4Γ + I model, Metropolis-coupled Markov chain Monte Carlo sampling was carried out with one cold and three heated chains starting from random starting trees and the program default prior probabilities on model parameters. The maxspe data were run for 3,000,000 generations (average standard deviation [SD] of split frequencies <0.0025), and the maxgen data were run for 1,000,000 generations (average SD of split frequencies <0.0000). For both data sets, samples of the Markov chain were taken every 100 generations giving a total sample of 30,000 trees (maxspe) or 10,000 trees (maxgen). Parameters were checked for stationarity with Tracer v1.4 (Rambaut and Drummond 2007), and the first 10,000 trees were discarded as burn-in. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled after the initial burn-in period.

Results

ESTs and Alignments from Baetis sp. and I. elegans

After trimming vectors, filtering for minimum length (<100 bp) and removal of low-quality sequences, we obtained 4,197 Baetis sp. sequences from the initial 4,225 clones. For I. elegans, we obtained 4,217 from 4,219 randomly sequenced clones. Clustering resulted in 3,035 contigs (635 contigs contain more than one EST, 2,400 singletons) for Baetis sp. and 3,194 (614 contigs contain more than one EST, 2,580 singletons) for I. elegans.

Based on the HaMStR approach (Ebersberger et al. 2009), we identified 436 orthologous sequences in Baetis sp. and 527 orthologous sequences in I. elegans.

Due to the limited number of assigned orthologs in each species, the data sets differed significantly in their represented species, genes, proportion of missing data within the taxa, and their overall sequence length. In the maxspe alignment, 15 species and 125 genes were represented with an alignment length of 31,643 aa and a proportion of missing data of 18%. The maxgen alignment maximized the represented genes (150) but reduced the taxa number to eight, had a sequence length of 42,541 aa, and a proportion of missing data of 11%. For an overview of represented genes in both data sets, see supplementary table S2 (Supplementary Material online).
Phylogenetic Analyses of Concatenated Alignments

After the exclusion of randomly similar aligned sections identified by ALISCORE (Misof B and Misof K 2009), the data set maxspe comprised 27,327 aa (initial 31,643 aa, ~14% randomly similar) and maxgen comprised 37,473 aa (initial 42,541 aa, ~12% randomly similar). The final alignments have been deposited at TREEBASE (http://www.treebase.org, study GenBank accession number S2456). Results of the hypotheses testing using heuristic search and incorporating topology constraints are summarized in table 1. Based on the constrained analyses, the Chiastomyaria scenario (Odonata, Ephemeroptera + Neoptera) is significantly supported by all tests (AU, KH, SH, WKH, and WSH) in the maxspe data sets, whereas the maxgen alignment could not significantly reject the Metapterygota theory in the WSH test (WSH = 0.062) using the 95% significance level.

We employed ML and BI analyses to construct phylogenetic trees from the maxspe alignment (fig. 2) and the maxgen alignment (fig. 3). In all trees, I. elegans (Odonata) represent—with high bootstrap support/posterior probability (maxspe: 100%/100%; maxgen: 100%/100)—the most basal winged insect specimens, supporting the Chiastomyaria theory. The topology generated from the maxspe alignment further supports the monophyly of Paraneoptera (Aphis gossypii, Macconellicoccus hirsutus; 98%/100%) and Holometabola (100%/100%), with a basal position of Hymenoptera within the Holometabola data set (fig. 2). The relationships within the Lepidoptera were not well supported in the ML (62-31%) and the BI (32-28%) analyses based on the maxspe data set.

We also note that the tree based on the maxgen alignment is a true subtree of the maxspe tree. This indicates that the results are robust with respect to the number of species and genes. To further evaluate the quality of fit for the chosen model of evolution, we performed the test developed by Goldman (1993). The results (see supplementary figs. S1 and S2, Supplementary Material online) support that the WAG model describes the data adequately.

Table 1

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Hypothesis</th>
<th>AU</th>
<th>KH</th>
<th>SH</th>
<th>WKH</th>
<th>WSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxgen</td>
<td>Palaeoptera</td>
<td>2 × 10^{-104}**</td>
<td>0.001*</td>
<td>0.001*</td>
<td>0.001*</td>
<td>0.001*</td>
</tr>
<tr>
<td></td>
<td>Metapterygota</td>
<td>0.052*</td>
<td>0.035*</td>
<td>0.039*</td>
<td>0.035*</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>Chiastomyaria</td>
<td>0.971</td>
<td>0.965</td>
<td>0.987</td>
<td>0.965</td>
<td>0.985</td>
</tr>
<tr>
<td>maxspe</td>
<td>Palaeoptera</td>
<td>7 × 10^{-50}***</td>
<td>0.002*</td>
<td>0.002*</td>
<td>0.001*</td>
<td>0.001*</td>
</tr>
<tr>
<td></td>
<td>Metapterygota</td>
<td>0.029*</td>
<td>0.025*</td>
<td>0.025*</td>
<td>0.025*</td>
<td>0.048*</td>
</tr>
<tr>
<td></td>
<td>Chiastomyaria</td>
<td>0.971</td>
<td>0.975</td>
<td>0.982</td>
<td>0.975</td>
<td>0.979</td>
</tr>
</tbody>
</table>

Fig. 2.—ML + BI topology of maxspe. Pterygote phylogenetic relationships based on 15 taxa and 125 genes data set (maxspe) showing a basal position of Odonata (Ischnura elegans), the monophyly of Paraneoptera and Holometabola. Branch lengths are from ML trees. Bootstrap support values of ML analysis and Bayesian posterior probabilities for each branch are indicated before and after a slash, respectively. Asterisk indicates 100% support value, hyphen indicates support value below 50%.
Phylogenetic Analyses of Single Alignments

Both data sets were scanned for individual genes represented in *Baetis* sp., *I. elegans*, and *Onychiurus arcticus*, as well as in at least one neopterous insect. In the *maxspe* alignment, we identified 39 genes and in the *maxgen* alignment 58 genes. Of these, 34 genes are present in both alignments. The function of these 63 genes was assessed through Blast against the eukaryotic orthologous groups (KOGs) database (http://biotec.icb.ufmg.br/K-EST/begin.html) and assigned to the four major KOG categories: 1) cellular processes and signaling, 2) information storage and processing, 3) metabolism, and 4) poorly characterized (table 2).

We performed extended ML tree analyses of the individual *maxspe* (total 39) and *maxgen* (total 58) alignments to investigate the support of the three phylogenetic hypotheses by the individual genes. The log likelihood for each topology was calculated using Tree-Puzzle 5.2 (Schmidt et al. 2002). The topologies were considered as supported by the individual gene alignments if the P-SH < 0.05 and if the ΔlogL: standard error (SE) ratio exceeded 0.5 (supplementary table S3, Supplementary Material online). In addition, for each gene alignment of the *maxspe* (supplementary table S4a, Supplementary Material online) and *maxgen* set (supplementary table S4b, Supplementary Material online), which included a sequence of *Baetis* sp., *I. elegans*, *O. arcticus*, and at least one neopterous insect, an ML tree with 100 bootstrap replicates was calculated using RAxML. Within *maxgen*, based on the P-SH value and the ΔlogL: SE ratio, two genes (lethal [2] tumorous imaginal discs and Helicase at 2SE) support the Metapterygota hypothesis and the gene Cysteine proteinase Cathepsin L (K-EST description) supports the Chiastomyaria hypothesis. The majority of the genes (55) represented in the *maxgen* set did not carry sufficient phylogenetic signal to distinguish between the three alternative topologies (supplementary table S3, Supplementary Material online). In addition, the bootstrap analyses for each gene alignment did not provide significant support (>95%) for a single phylogenetic hypothesis (supplementary table S4a, Supplementary Material online). To increase phylogenetic signal, the genes of the *maxgen* data set were concatenated according to their KOG category and subjected to ML tree analyses using the same methods as in the individual gene analyses. Table 3 summarizes the support for the three phylogenetic hypotheses as recoded for analyses based on the functional classification using the statistical methods. The proteins involved in cellular processes and signaling (concatenated = 5,285 aa) gave the strongest support for the Chiastomyaria hypothesis and rejected significantly both other topologies. The proteins contained in the metabolism category (concatenated = 3,143 aa) also favor the Chiastomyaria hypothesis but did not significantly reject the Metapterygota hypothesis. Proteins classified as information storage and processing proteins (concatenated = 4,697 aa) favor the Metapterygota hypothesis but did not reject the Chiastomyaria hypothesis. The poorly characterized proteins (1,179 aa) identified the Metapterygota topology as the best but again did not reject the remaining hypothesis.

Some of the individual *maxspe* alignments, which were also subjected to extended ML tree analysis using Tree-Puzzle and RAxML, provide significant support for one of the phylogenetic hypotheses (see supplementary tables S3 and S4a, Supplementary Material online). To increase the phylogenetic signal, we also concatenated the individual *maxspe* alignments based on their KOG category assignment (cellular processes and signaling [3,511 aa], information storage and processing [4,245 aa], metabolism [1,551 aa], and poorly characterized [329 aa]). Three of the four KOG category–derived *maxspe* alignments identified the Chiastomyaria phylogeny as the best ML tree, but the two alternative topologies could not be rejected by the
<table>
<thead>
<tr>
<th>KOG Category</th>
<th>ID Number</th>
<th>FlyBaseID</th>
<th>Gene Name (FlyBase)</th>
<th>maxspe</th>
<th>maxgen</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Cellular processes and signaling</td>
<td>6,936</td>
<td>FBgn0038166</td>
<td>CG9588</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7,538</td>
<td>FBgn0034709</td>
<td>CG3074</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,640</td>
<td>FBgn0015282</td>
<td>Proteasome 26S subunit 4 ATPase</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,073</td>
<td>FBgn0023174</td>
<td>Proteasome β2 subunit</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,075</td>
<td>FBgn0003150</td>
<td>Proteasome 29-kD subunit</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,671</td>
<td>FBgn0033663</td>
<td>ERp60</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,489</td>
<td>FBgn0062174</td>
<td>Lethal (2) tumorous imaginal discs</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,547</td>
<td>FBgn0010638</td>
<td>Sec61β</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,032</td>
<td>FBgn0010226</td>
<td>Glutathione S-transferase S1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,784</td>
<td>FBgn0011127</td>
<td>Effete</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,782</td>
<td>FBgn0010602</td>
<td>Lesswright</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,616</td>
<td>FBgn0037756</td>
<td>CG8507</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,827</td>
<td>FBgn0025637</td>
<td>skpA</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,864</td>
<td>FBgn0036928</td>
<td>Translocase of outer membrane 20</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,902</td>
<td>FBgn0037231</td>
<td>CG9779</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,323</td>
<td>FBgn0024833</td>
<td>AP-47</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,169</td>
<td>FBgn0021814</td>
<td>Vps28</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,720</td>
<td>FBgn0025700</td>
<td>CG5885</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,562</td>
<td>FBgn0028985</td>
<td>Serine protease inhibitor 4</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,339</td>
<td>FBgn0011760</td>
<td>Cutup</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,414</td>
<td>FBgn0052672</td>
<td>Autophagy-specific gene 8a</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>(2) Information storage and processing</td>
<td>9,511</td>
<td>FBgn0014189</td>
<td>Helicase at 25E</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,970</td>
<td>FBgn0011997</td>
<td>Histone H2A variant</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,512</td>
<td>FBgn0037346</td>
<td>Extra bases</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,671</td>
<td>FBgn0029897</td>
<td>Ribosomal protein L17</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,790</td>
<td>FBgn0001942</td>
<td>Eukaryotic initiation factor 4a</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,906</td>
<td>FBgn0034967</td>
<td>Eukaryotic initiation factor 5A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,927</td>
<td>FBgn0037351</td>
<td>Ribosomal protein L13A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,007</td>
<td>FBgn0010265</td>
<td>Ribosomal protein S13</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,098</td>
<td>FBgn0036213</td>
<td>Ribosomal protein L10Ab</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,316</td>
<td>FBgn0034743</td>
<td>Ribosomal protein S16</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,606</td>
<td>FBgn0005593</td>
<td>Ribosomal protein L7</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,883</td>
<td>FBgn0039713</td>
<td>Ribosomal protein S8</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,950</td>
<td>FBgn0034751</td>
<td>Ribosomal protein S24</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,013</td>
<td>FBgn0002590</td>
<td>Ribosomal protein S5a</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,023</td>
<td>FBgn0010409</td>
<td>Ribosomal protein L18A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,456</td>
<td>FBgn0034138</td>
<td>Ribosomal protein S15</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,732</td>
<td>FBgn0064225</td>
<td>Ribosomal protein L5</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,997</td>
<td>FBgn0039129</td>
<td>Ribosomal protein S19b</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,404</td>
<td>FBgn0031980</td>
<td>Ribosomal protein L36A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,821</td>
<td>FBgn0036825</td>
<td>Ribosomal protein L26</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,715</td>
<td>FBgn0024558</td>
<td>Diphthamide methyltransferase</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,590</td>
<td>FBgn0028737</td>
<td>Elongation factor LF</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,383</td>
<td>FBgn0023211</td>
<td>Elongin C</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,771</td>
<td>FBgn0023212</td>
<td>Elongin B</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>(3) Metabolism</td>
<td>6,637</td>
<td>FBgn0014028</td>
<td>Succinate dehydrogenase B</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,384</td>
<td>FBgn0011361</td>
<td>Mitochondrial acyl carrier protein 1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,813</td>
<td>FBgn0031436</td>
<td>CG3214</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,569</td>
<td>FBgn0028662</td>
<td>VhaPPA1-1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,434</td>
<td>FBgn0039697</td>
<td>CG7834</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,214</td>
<td>FBgn0000116</td>
<td>Arginine kinase</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,594</td>
<td>FBgn0028014</td>
<td>CG4169</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6,958</td>
<td>FBgn0036580</td>
<td>PDCD-5</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,095</td>
<td>FBgn0028833</td>
<td>Dak1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>9,007</td>
<td>FBgn0025083</td>
<td>Deoxyuridine triphosphatase</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,631</td>
<td>FBgn0032192</td>
<td>CG5731</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,076</td>
<td>FBgn0033879</td>
<td>CG6543</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,736</td>
<td>FBgn0035528</td>
<td>CG15012</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>7,742</td>
<td>FBgn0038739</td>
<td>CG4686</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>8,092</td>
<td>FBgn0030724</td>
<td>Nipsnap</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Table 2**

**Genes Shared between *Baetis* sp., *Ischnura elegans*, and *Onychiurus arcticus*, as well as At Least One Neopterous Insect**

**KOG Category**

(1) Cellular processes and signaling

(2) Information storage and processing

(3) Metabolism

(4) Poorly characterized

**NOTE.**—These genes were assembled in the four major KOG categories: 1) cellular processes and signaling, 2) information storage and processing, 3) metabolism, and 4) poorly characterized. ID number—the numerical identifier assigned to the gene during the HaMStR process; FlyBaseID/gene name—the corresponding ID number/gene name of the *Drosophila melanogaster* genome database (http://flybase.org/). maxspe/maxgen—genes represented in the alignments. These genes were also selected for the extended ML analyses of individual alignments.
proteins involved in information storage and processing + metabolism, whereas the genes involved in cellular processes and signaling significantly support the Chiastomyaria theory. Proteins categorized as poorly characterized identified the Palaeoptera topology as the best tree but not significantly (summarized as table 3).

### Discussion

The question of the first winged insect order has been dominated by the analyses of morphological characters and nuclear RNA data (18S and 28S). Recently, Zhang et al. (2008) published the first mitochondrial genome of an Ephemeroptera. The analysis used the mitogenomic approach and supported the Metapterygota hypothesis. Despite numerous studies concerning the phylogenetic relationships at the base of pterygotes, the so-called Palaeoptera problem is still not solved and results are often conflicting. A combined analysis of nuclear rRNA (18S and 28S) and 275 morphological characters supported the Metapterygota hypothesis (Wheeler et al. 2001) as did a combined analysis of 18S + 28S rRNA, the protein-coding gene Histone 3, and morphology data (Ogden and Whiting 2003). This hypothesis is supported by some diagnostic morphological characters connecting Ephemeroptera with the apterygote hexapods, such as molting, muscle structure in the tracheal system, and the caudal filament (Kristensen 1991). However, different analyses of nuclear rRNA data by different authors support each of the three phylogenetic hypotheses depending on the phylogenetic inference method used, for example, combined 18S and 28S supports the Metapterygota hypothesis (Wheeler et al. 2001; Ogden and Whiting 2003), the Palaeoptera hypothesis (Hovmöller et al. 2002), and the Chiastomyaria hypothesis (Mallatt and Giribet 2006; von Reumont et al. 2009). The longest standing hypothesis and the traditional textbook scenario based on morphological characters is the Palaeoptera hypothesis. It is supported by the inability of the Ephemeroptera and Odonata to fold their wings over the abdomen (Hennig 1969; Kukalova-Peck 1991), the intercalary veins in the wings, the fusion of the galea and lacinia in the larval maxillae, and the aquatic larvae (Hennig 1981). Kjer et al. (2006) also supported this hypothesis using 9 genes and 170 morphological characters. However, a strong argument for the third hypothesis—the Chiastomyaria hypothesis—is the indirect sperm transfer mechanism linking the Odonata to the apterygote insects (Boudreaux 1979) and the direct flight muscles that are a unique character of Odonata. This phylogenetic hypothesis is further supported by several molecular studies (Kjer 2004; Yoshizawa and Johnson 2005; Misof et al. 2007).

All studies clearly illustrate that basal pterygote divergence is difficult to unveil, despite the use of various morphological characters and molecular markers. One major problem is certainly the fast evolution of the pterygotes and the enormous diversity within this group. Furthermore, the preserved ancient characters in some taxa and the rate heterogeneity among orders lead to confusion among phylogeneticists. For example, Kjer et al. (2006) observed excessive substitution rate acceleration for Diptera and Diplura, whereas Odonata and Mantodea seem to almost “standstill.” Finding appropriate molecular markers with phylogenetically informative sites tracking the narrow window, within which the divergence and origin of winged insects took place, is the major challenge. In this study, we included two crucial new basal winged insect EST data sets (representing the Odonata and Ephemeroptera), adopted a multigene approach, and evaluated the support of different classes of functional protein-coding genes for each of the three hypotheses.

Protein-coding sequences obtained by EST sequencing represent a valuable and relatively inexpensive possibility for resolving long outstanding deep phylogenetic relationships. The conserved nature of the housekeeping genes makes studies of divergences that took place millions of years ago possible. Thus, phylogenetic trees inferred from multigene approaches using ESTs have become a popular method to resolve long outstanding questions in deep metazoan relationships. Dunn et al. (2008), for example, improve the resolution of the animal tree of life using a concatenated alignment of 150 genes, Philippe et al. (2004) concatenated 129 orthologous proteins for eukaryotic species, and Savard et al. (2006) assembled 185 genes to resolve the radiation of Holometabolous insects. The advantages of a multigene approach instead of a single gene or few genes are numerous. Rokas, Williams, et al. (2003) pointed out that the biological process of a gene as influenced by natural selection or genetic drift may cause the history of the genes under analysis to obscure the history of the taxa. Issues such as gene duplication and lineage sorting may contribute to varying degrees of discordance between gene tree and species tree. Therefore,

### Table 3
ML Support for the Three Different Phylogenetic Hypotheses of the Concatenated Alignments Based On Their KOG Category

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Hypothesis</th>
<th>Cellular Processes and Signaling</th>
<th>Information Storage and Processing</th>
<th>Metabolism</th>
<th>Poorly Characterized</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>maxgen</em></td>
<td>Palaeoptera</td>
<td>6.61 &lt;0.0000***</td>
<td>2.13 0.0570</td>
<td>0.87 0.2040</td>
<td>1.06 0.2040</td>
</tr>
<tr>
<td></td>
<td>Metapterygota</td>
<td>6.64 &lt;0.0000***</td>
<td>[31382.03] 1.00</td>
<td>2.05 0.0290</td>
<td>[19394.10] 0.0290</td>
</tr>
<tr>
<td></td>
<td>Chiastomyaria</td>
<td>[47438.79] 1.00</td>
<td>[44772.80] 1.00</td>
<td>[19394.10] 0.0290</td>
<td></td>
</tr>
<tr>
<td><em>maxspe</em></td>
<td>Palaeoptera</td>
<td>6.79 &lt;0.0000***</td>
<td>1.64 0.0910</td>
<td>0.74 0.2650</td>
<td>[3228.31] 0.2650</td>
</tr>
<tr>
<td></td>
<td>Metapterygota</td>
<td>6.77 &lt;0.0000***</td>
<td>0.37 0.4810</td>
<td>1.41 0.1160</td>
<td>1.04 0.1160</td>
</tr>
<tr>
<td></td>
<td>Chiastomyaria</td>
<td>[46215.35] 1.00</td>
<td>[43602.25] 1.00</td>
<td>[22195.41] 1.00</td>
<td></td>
</tr>
</tbody>
</table>

Note: —The favored topology of each KOG category is indicated in bold. The support is expressed as the AlogL: SE and the P-SH value. The ~logL. value of the best tree is written in square brackets.
conflicting topologies are often seen in analyses of a single or small numbers of concatenated genes. Furthermore, the use of one or a few genes is known to be insufficient for the resolution of many clades (Bapteste et al. 2002; Rokas, King, et al. 2003; Rokas, Williams, et al. 2003), whereas larger amounts of data and the increasing number of phylogenetic informative positions robustly resolve the topology (Philippe et al. 2004). However, is a multigene approach really a panacea for the accurate resolution of a species tree? A study by Gadagkar et al. (2005) indicates that this may not be the case, by showing that weak phylogenetic signals can be substantially reinforced when sequences are concatenated, but in the worst case, it can also enhance support for the erroneous inferences, leading to very high bootstrap support for incorrect clades. In other words, the multigene approach does not necessarily lead to the correct topology because adding of new genes does not increase the accuracy of the topology in the presence of a bias. Various studies have shown that the consistency of tree reconstruction in phylogenomic studies is sensitive to the model of sequence evolution (Phillips et al. 2004; Jeffroy et al. 2006) and to taxon sampling (Hillis et al. 2003; Brinkmann et al. 2005), both potential sources of LBA artifacts. Subsequently, the detection and avoidance of LBA artifacts remain the most important challenge for phylogenomic studies. One strategy to reduce the impact of systematic bias would be to apply probabilistic methods that take into account variable evolutionary rates over sites and lineages (Kolaczkowski and Thornton 2004; Brinkmann et al. 2005). Unfortunately, no current model covers the full complexity of biological history that can minimize the inconsistency of methods caused by model misspecification (Steel 2005).

In this study, we have attempted to identify the impact of systematic bias in our phylogenetic analyses by applying suitable methods of analysis to better match the data and did not detect any severe model violations (see supplementary figs. S1–S4, Supplementary Material online). Adequate taxon sampling remains the other crucial factor in phylogenomic studies to avoid LBA artifacts. Increasing the number of ingroup taxa from 7 (maxgen) to 14 (maxspe) resulted in a congruent topology and support for the Chiastomyaria hypothesis, that is, a basal position of Odonata. However, given the existing data, we are not in the position to significantly enlarge taxon sampling. At this time, the Chiastomyaria hypothesis is well supported, but we are aware that possible pitfalls (LBA, wrong model of sequence evolution, gene sampling) exist. Thus, future extended analyses are necessary to finally confirm the Chiastomyaria hypothesis.

On the other hand, not only is the phylogenomic methodology or taxon sampling important but the genes/proteins to which it is applied are also of relevance. The evolutionary history of the genes that compose the data sets may have a direct impact on the reconstructed phylogeny (Comas et al. 2007). The phylogenetic signal of a gene is likely to be related to its evolutionary constraint, and it has been suggested that a polytomy can be resolved by using genes that evolve at the optimal rate in the relevant timescale (Townsend 2007).

We therefore assessed the biological function of the represented genes and concatenated them according to their functional classification with the assumption that they harbor the same evolutionary history along the branches of the organismal phylogeny. It has been known that different evolutionary signals are a result of the different evolutionary processes that act upon the genes and that the functional role of these genes in the cell is important for the phylogenetic signal they carry (Graur and Li 2000).

The statistical tests of concatenated alignments based on their functional classification showed that proteins belonging to the cellular processes and signaling category seem to harbor the strongest phylogenetic signal for resolving deep phylogenetic relationships. Our results are congruent with a phylogenetic study of the fungal kingdom (Kuramae et al. 2007). These authors evaluated phylogenetically informative proteins for the fungal Tree of Life and identified proteins involved in cellular processes and signaling as phylogenetically more informative than the others.

Nevertheless, the large dataset based on KOG categories (maxgen: cellular processes and signaling = 5,285 aa, information storage and processing = 4,697 aa, metabolism = 3,143 aa, and poorly recognized proteins = 1,179 aa; maxspe: cellular processes and signaling = 3,511 aa, information storage and processing = 4,245 aa, metabolism = 1,551 aa, and poorly recognized proteins = 329 aa) gave in the majority of analyses no strong statistical support for any one hypothesis. There are several explanations for this observation. First of all, multiple substitutions at the same positions are expected to be frequent because the speciation event occurred millions of years ago. The saturation of the molecular markers will certainly reduce the phylogenetic signal and consequently the resolution. To investigate this, we conducted ML analyses for each protein separately using Tree-Puzzle (WAG + 4I + 1) and RAxML (PROTMIXWAG). As expected, due to the limited number of alignment positions, the analyses from the individual alignments have shown that one gene did not harbor enough phylogenetic signal to unequivocally resolve the Palaeooptera problem. Although the conserved nature of housekeeping genes is beneficial to track Mesozoic divergences, the phylogenetic content of single genes is too low, whereas concatenation seems to compensate for this fact.

It appears that the ancient rapid radiation that took place with the transition from nonwinged to winged insects represents one of the major obstacles for insect systematics. As we have shown for one of the major questions in insect phylogeny, molecular phylogenetics may overcome this hurdle by closing the gaps of genetic information from key orders, carefully applying multigene approaches, and assessing the data quality.

Supplementary Material

Supplementary tables S1–S4 and figures S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]) special
priority program “Deep Metazoan Phylogeny” SP 1174
grant given to H.H. (DFG HA 1947/5), A.v.H. (DFG
HA 1628/9), and the Boehringer Ingelheim Fonds given
to S. Simon that supported the Workshop on Molecular
Evolution 2008, Marine Biological Laboratory, Woods
Hole, MA, USA. A.v.H. and S. Strauss would also like to
thank the WWTF for generous funding. S. Simon would
like to express her gratitude to Michael P. Cummings, Ste-
ven Thompson, and Akito Y. Kawahara for helpful sugges-
tions concerning data analyses during the MBL Workshop
in Woods Hole. We thank Sara Khadjeh for providing
tations concerning data analyses during the MBL Workshop
ven Thompson, and Akito Y. Kawahara for helpful sugges-
tions. Special thanks go also to Ingo Ebersberger (Center of Integrative
Bioinformatics Vienna, Vienna) for providing a prerelease
of the HaMSTR tool and the PERL script to calculate the
amount of missing data, Danielle de Jong for linguistic help, and Bernd Schierwater for his helpful comments
on an earlier version of the manuscript. We also thank
Associate Editor Barbara Holland and two anonymous
reviewers for providing constructive comments that greatly
improved the manuscript.

Literature Cited

best-fit models of protein evolution. Bioinformatics. 21:
2104–2105.

Parallel Metropolis coupled Markov chain Monte Carlo for

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z,
Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-
BLAST: a new generation of protein database search programs.

New

Highly divergent amoebae: Dictyostelium, Entamoeba, and
results concerning the morphology of the most ancient
dragonflies (Insecta: Odonatoptera) from the Namurian of

Belfiore NM, Liu L, Moritz C. 2008. Multilocus phylogenetics of
mammals (Arthropoda): evolutionary patterns and

Genome Res. 14:988–995.


Brinkmann H, van der Giezen M, Zhou Y, Poncelin de
long-branch attraction artefacts in deep eukaryotic phyloge-

Chou HH, Holmes MH. 2001. DNA sequence quality trimming

signal and functional categories in Proteobacteria genomes.

Cotton JA, Wilkinson M. 2009. Supertrees join the mainstream of

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the
reconstruction of the tree of life. Nat Rev Genet. 6:
361–375.

Broad phylogenomic sampling improves resolution of the

hidden Markov model based search for orthologs in ESTs.

trees without concatenation. Proc Natl Acad Sci USA. 104:
5936–5941.

Engel MS, Grimaldi DA. 2004. New light shed on the oldest

Felsenstein J. 1978. Cases in which parsimony or compatibility
methods will be positively misleading. Syst Zool. 27:401–410.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species
phylogenies from multiple genes: concatenated sequence tree
304:64–74.


Gaunt MW, Miles MA. 2002. An insect molecular clock dates
the origin of the insects and accords with palaeontological and

Goldman N. 1993. Statistical tests of models of DNA substitu-


Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of
the heterogeneity of substitution rate among nucleotide sites.

identification and analysis of transcripts differentially regu-
lated in fracture healing by EST sequencing in the domestic
sheep. BMC Genomics. 7:172.

Hennig W. 1969. Die Stammesgeschichte der Insekten. Frankfurt
(Germany): Kramer. p. 436.


taxon sampling a problem for phylogenetic inference? Syst

problem: basal pterygote phylogeny inferred from 18S and
28S rDNA sequences. Cladistics. 18:313–323.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference

Hughes J, Longhorn SJ, Papadopoulou A, Theodorides K, de
Riva A, Mejia-Chang M, Foster PG, Vogler AP. 2006. Dense
taxonomie EST sampling and its applications for molecular
23:268–278.


Zwickl D. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Austin (TX): The University of Texas.


Barbara Holland, Associate Editor

Accepted August 18, 2009