Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta)

Barbara J. Sharanowski [a,*], Barbara Robbertse [b], John Walker [c], S. Randal Voss [c], Ryan Yoder [b], Joseph Spatafora [b], Michael J. Sharkey [a]

[a] Department of Entomology, University of Kentucky, S-225 Agricultural Science Center North, Lexington, KY 40546–0091, USA
[b] The Center for Genome Research and Biocomputing, 3021 Agricultural and Life Sciences Building, Oregon State University, Corvallis, OR 97331-7303, USA
[c] Department of Biology, University of Kentucky, Chandler Medical Center, B453 Biomedical & Biological Sciences Research Building, 741 S. Limestone Street, Lexington, KY 40536-050, USA

ARTICLE INFO

ABSTRACT

Hymenoptera is one of the most diverse groups of animals on the planet and have vital importance for ecosystem function as pollinators and parasitoids. Higher-level relationships among Hymenoptera have been notoriously difficult to resolve with both morphological and traditional molecular approaches. Here we examined the utility of expressed sequence tags for resolving relationships among hymenopteran superfamilies. Transcripts were assembled for 6 disparate Hymenopteran taxa with additional sequences added from public databases for a final dataset of 24 genes for 16 taxa and over 10 kb of sequence data. The concatenated dataset recovered a robust and well-supported topology demonstrating the monophyly of Holometabola, Hymenoptera, Apocrita, Aculeata, Ichneumonoidea, and a sister relationship between the two most closely related proctotrupomorphs in the dataset (Cynipoidea + Proctotrupoidea). The data strongly supported a sister relationship between Aculeata and Proctotrupomorpha, contrary to previously proposed hypotheses. Additionally there was strong evidence indicating Ichneumonoidea as sister to Aculeata + Proctotrupomorpha. These relationships were robust to missing data, nucleotide composition biases, low taxonomic sampling, and conflicting signal across gene trees. There was also strong evidence indicating that Chalcidoidea is not contained within Proctotrupomorpha.

1. Introduction

Recently, researchers have demonstrated the power of utilizing genomic information for phylogenetic reconstruction (Dunn et al., 2008; Philippe et al., 2005; Rokas et al., 2003; Savard et al., 2006). Particularly, expressed sequence tags (ESTs), which are fragments of coding sequence, offer an abundant and efficient source of new genetic markers for phylogenetic analysis (Hughes et al., 2006). Utilizing ESTs also allows for amplification of a wider range of taxa than just those species involved in whole genome sequencing projects. Additionally, datasets based on ESTs utilize significantly more genetic information than traditional polymerase chain reaction (PCR) approaches. As the number of independent molecular markers increases, gene trees can converge upon a more accurate species tree (Rokas et al., 2003; Savard et al., 2006, but see

Degnan and Rosenberg, 2006). However, this phylogenomic approach is often weakened by limited taxon sampling, which may increase systematic error (Baurain et al., 2007; Dávalos and Perkins, 2008; Zwickl and Hillis, 2002).

The main purpose of this paper is to test the utility of using ESTs for phylogenetic analysis of Hymenoptera at the superfamily level. Hymenoptera (Insecta), includes the bees, ants, and parasitoid wasps and constitutes one of the most important and diverse group of organisms on earth from both an anthropogenic and environmental perspective (Austin and Dowton, 2000; Gauld and Bolton, 1988; Whitfield, 1998). Members of Hymenoptera are invaluable insects to humans, working as efficient parasitoids of destructive pests, as important pollinators of plants, and as major contributors to ecosystem function. Unfortunately, there is little understanding of the phylogenetic relationships among superfamilies, particularly among the highly diverse parasitic lineages. Several studies have attempted to resolve higher-level Hymenopteran relationships using morphological data (Königsmann, 1976, 1978a,b; Rasnitsyn, 1988; Ronquist et al., 1999; Vilhelmsen et al., 2010), molecular data (Castro and Dowton, 2006, 2007; Dowton and Austin, 1994; Dowton et al., 1997), or a combination of both (Carpenter and Wheeler, 1999; Dowton and Austin, 2001).

Morphological datasets have been hampered by convergent homoplastic characters typical among parasitoids, as unrelated organisms may possess the same phenotypic adaptations for parasitizing similar hosts. Molecular datasets have thus far been restricted to mitochondrial and ribosomal DNA markers that are easy to amplify across a wide range of taxa. While taxonomic sampling has been considerable in most molecular datasets produced to date, the limited number of genetic loci has failed to provide robust resolution at the level of superfamily. Thus, even after almost 40 years of study using phylogenetic techniques, there is still a great deal of uncertainty regarding patterns of Hymenopteran evolution (Sharkey, 2007). This lack of knowledge prevents understanding of the mode and pattern of evolutionary traits, such as the evolution of parasitism strategies, social behavior, complex venoms, and polydna viruses (Whitfield, 1998; Whitfield et al., 2003).

Here we test the utility of using ESTs for phylogenetic analysis of Hymenoptera at the superfamily level. The dataset includes 10 hymenopteran taxa, with six of these newly sequenced for representative transcripts. Taxon sampling includes representatives of superfamilies that have been historically unresolved. This paper presents the first attempt to reconstruct hymenopteran evolutionary relationships utilizing nuclear protein coding genes and a phylogenomics approach.

2. Taxonomic background

Hymenoptera has traditionally been divided into two suborders, Symphyta, or sawflies, and Apocrita, or wasp-waisted hymenopterans. While the monophyly of the Apocrita has long been recognized (Königsmann, 1978a; Rasnitsyn, 1988; Ronquist et al., 1999), Symphyta is now acknowledged as a paraphyletic basal grade (Schulmeister et al., 2002; Vilhelmsen, 2001). The Apocrita has further been subdivided into two groups: the Aculeata, containing the bees, ants, and stinging wasps; and the Parasitica, of

which most of the members are parasitoids of insects and arachnids. The Parasitica, containing the majority of the diversity of the order, has been the least understood group and is likely paraphyletic with respect to Aculeata (Brothers, 1975). Rather than utilizing these two unnatural but traditional subdivisions, Rasnitsyn (1988) proposed a new infraorder system for the extant apocritan lineages (=suborder Vespina), including Orussomorpha, Evaniomorpha, Proctotrupomorpha, Ichneumonomorpha, and Vespomorpha (more traditionally known as Aculeata). Although Rasnitsyn suggested the Orussoidea were contained within Apocrita, other researchers suggest that the parasitic Orussoidea is the sister group to Apocrita (Sharkey, 2007; Vilhelmsen, 2003). Similar to Brothers (1975), Rasnitsyn proposed a sister relationship between Ichneumonomorpha and Vespomorpha. Additionally, he suggested that the Evaniomorpha and Proctotrupomorpha are sister groups. These relationships are summarized in Fig. 1A.

The proposed Proctotrupomorpha (including Cynipoidea, Proctotrupoidea, Platygastroidea, and Chalcidoidea) was a novel hypothesis differing from his earlier work (Rasnitsyn, 1980) that had placed these superfamilies with the Ichneumonoidea. Interestingly, when Ronquist et al. (1999) reanalyzed Rasnitsyn's (1988) morphological data using cladistic techniques, the relationships proposed by Rasnitsyn were not recovered, with most conflicting resolution attributable to reductional characters (see Sharkey and Roy, 2002). However, Rasnitsyn's (1988) proposed classification of the Hymenoptera has probably been the most widely accepted, or at least tested hypothesis (for a full review, see Sharkey, 2007; and Whitfield, 1992). Unfortunately, the relationships between most apocritan lineages have lacked stability across independent datasets, leaving ample doubt about the validity of these associations.

Dowton and Austin (1994) performed one the first molecular analyses of Hymenoptera based on one mitochondrial gene (16S rRNA). While most relationships were not resolved, they did recover a sister relationship between Ichneumonoidea and Aculeata, as proposed by Rasnitsyn (1988), albeit with very low nodal
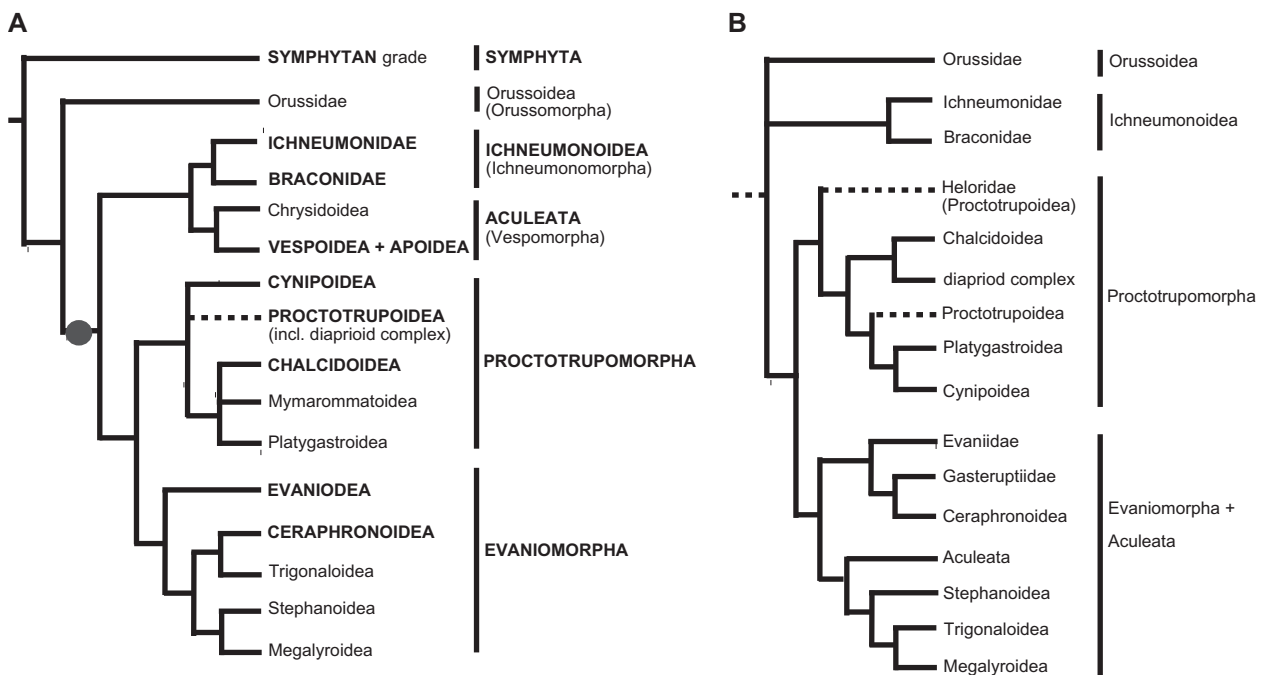
Fig. 1. (A) Summarized tree demonstrating relationships among the major extant lineages within Hymenoptera. This is a modified diagram from Sharkey (2007) based on hypotheses proposed by Rasnitsyn (1988) from morphological data. The node indicated with a gray circle represents Apocrita. All taxa typed in uppercase are represented in the current study. (B) Summarized tree demonstrating relationships among the major extant lineages within Hymenoptera based on molecular data (28S, 18S, 16S, CO1) analyzed by Castro and Dowton (2006), (cf. Fig 2). Dashed lines indicate paraphyly.

support. Additionally they recovered a clade consistent with Rasnitsyn's (1988) Proctotrupomorpha, but again with little support. Carpenter and Wheeler (1999) performed a preliminary analysis of 36 Hymenopteran taxa for three genes (18S, 28S, and two regions of COI) and included the morphological dataset of Ronquist et al. (1999). While the combined analyses recovered a monophyletic Apocrita and Aculeata, all other clades demonstrated odd paraphyletic relationships. Dowton and Austin (2001) expanded their dataset in 2001 to include three genes (28S rRNA, 16S rRNA, and COI), 87 taxa and the morphological dataset from Ronquist (1999). They performed multiple analyses under variable partition and weighting schemes, but unfortunately the dataset was sensitive to analytical technique and the inclusion of morphology. Under at least one model, Dowton and Austin (2001, Fig. 5, p. 98) recovered a sister relationship between the Ichneumonoidea and Aculeata, as well as a monophyletic Proctotrupomorpha, but again these clades had relatively weak support.

More recently, Castro and Dowton (2006) employed Bayesian and Parsimony analyses on the Dowton and Austin (2001) dataset with the addition of 18S rRNA sequences (summarized in Fig. 1B). Unfortunately, several relationships were sensitive to outgroup selection, method of analysis, and gene inclusion. The phylogenetic position of Ceraphronoidea was typically recovered within Evaniomorpha or as a basal clade. Castro and Dowton (2006) recovered a monophyletic Proctotrupomorpha in most analyses with variable levels of support. However, the placement of Chalcidoidea within the Proctotrupomorpha was weakly supported in some analyses. In contrast to previous studies, there was consistent and strong support for Chalcidoidea as sister to the diaprioid complex (Diapriidae + Monomachidae + Maamingidae). Additionally, they typically recovered aculeates in a clade within Evaniomorpha (Fig. 1B), rather than sister to the Ichneumonoidea.

Molecular analyses of hymenopteran relationships have never incorporated nuclear protein coding genes, and this remains a potent source of genetic information that may be able to resolve relationships among apocritan superfamilies. Here, we test these relationships using expressed sequence tags as a source of molecular characters for a small subset of hymenopteran taxa representing 8 of the 15 extant apocritan superfamilies as recognized by Sharkey (2007). Obviously the 10 hymenopteran taxa utilized here do not represent a comprehensive sample of the taxonomic diversity within the order. However, this approach contrasts with the higher taxonomic, but low genetic sampling of previous analyses. Rather, a relatively large number of independent nuclear loci are utilized for a small number of taxa. Even with the low taxonomic sampling, it is possible to test the relationships proposed by Rasnitsyn (1988) and variably supported with molecular data, including: the monophyly of the Proctotrupomorpha; the sister group to Aculeata, and to a limited extent, the placement of Ceraphronoidea with respect to Evaniomorpha.

3. Materials and methods

3.1. Insect specimens

The extraction of RNA necessary for developing cDNA libraries requires extremely fresh and properly preserved specimens. The main motivation for taxon selection was to sample specimens that represented apocritan superfamilies that have been historically unresolved. In particular, attempts were made to obtain representative taxa from at least one symphytan and the following apocritan superfamilies: Ichneumonoidea, Proctotrupoidea, Ceraphronoidea, Evanioidea, Diaprioidea, and Cynipoidea. However, taxon selection was limited by the availability of extremely fresh material. Where possible, organisms were obtained from

established colonies. Additional material was obtained by collecting live material from the field, although it was not always possible to obtain multiple specimens for extraction or to establish exact identifications due to the limited number of specimens and the need to keep available specimens fresh while taxonomically identifying the organisms.

Of the six species of Hymenoptera sequenced for this experiment, two were obtained from existing colonies from colleagues as follows: the symphytan, *Neodiprion sertifer* (Hymenoptera: Tenthredinoidea: Diprionidae, (10 males, 10 females, Catherine Linnen, Harvard University); and *Campoletis sonorensis* (Hymenoptera: Ichneumonoidea: Ichneumonidae) (10 males, 10 females, Bruce Webb, University of Kentucky). Additionally, the diapriid, *Trichopria nigra* (Hymenoptera: Diaprioidea: Diapriidae), (10 males, 10 females, Kimberly Ferrero, University of Florida) was extracted for RNA but the quality was insufficient for cDNA library construction. The other four apocritan specimens were collected in Kentucky by the author (BJS) with a sweep net, including: *Pelecinus polyturator* (Hymenoptera: Proctotrupoidea: Pelecinidae) (2 females); *Pristaulacus strangliae* (Hymenoptera: Evanioidea: Aulacidae) (3 females); an unidentified ceraphronid (Hymenoptera: Ceraphronidae) (1 female); and an unidentified eucoiliine (Hymenoptera: Figitidae) (2 females). Specimens were stored whole at −80 °C until used. Table 1 lists all taxa in the analyses, including those whose sequences were mined from public databases, and the higher taxonomic names that are employed in all phylogenetic figures. Hymenopteran sequences mined for taxa from public databases were chosen based on availability. Outgroup sequences were chosen based on availability with an attempt to sample a broad range of taxa in which the relationships among outgroups have been well supported in other datasets. Additionally, annotated model genomes were utilized where possible to enhance the ability to determine orthology among loci.

3.2. RNA extraction and construction of cDNA libraries

Total RNA was extracted from all available specimens using TRIzol reagent (Invitrogen) (Chomczynski and Sacchi, 1987) according to the manufacturer's instructions and further cleaned using the RNeasy Mini Kit (Qiagen). The integrity of RNA of each species was analyzed on denaturing formaldehyde/agarose gel and quantified in a spectrometer to ensure a minimum of 50 ng starting material in a maximum of 3 μL. Additionally, RNA quantification and integrity assessments were performed on an Agilent 2100 bioanalyzer at the University of Kentucky MicroArray Core Facility.

Libraries were constructed using SMART™ cDNA Library Construction kit (Protocol PT3000-1, CLONTECH Laboratories), using the long-distance PCR method (Barnes, 1994; Chenchik et al., 1998). First strand cDNA synthesis was achieved using 1–3 μL of sample (0.05–1.0 μg total RNA), 20 units of Superscript II reverse transcriptase (Life Technologies), 1.2 μM SMART IV Oligonucleotide (5′-AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT ACG GCC GGG-3′), 1.2 μM CDS III/3′ PCR primer (5′-ATT CTA GAG GCC GAG GCG GCC GAC ATG-d(T)$_{30}$ (A/G/C)N-3′), 1 μM dNTP, 2 μM dithiothreitol (DTT), 1X buffer (50 mM Tris (pH 8.3), 6 mM MgCl$_2$, and 75 mM KCl) to a total volume of 10μL. Amplification of cDNA by PCR was performed in a GeneAmp 480 thermocycler using 5′PCR Primer (5′-AAG CAG TGG TAT CAA CGC AGA GT-3′) and CDS III/3′ PCR primer with the Advantage PCR kit (CLONTECH Laboratories) following the manufacturer's instructions. Thermocycler conditions were as follows: 1 min at 95 °C followed by 18–24 cycles of 15 s at 95 °C and 6 min at 68 °C. Subsequently, DNA polymerase activity was inactivated with proteinase K (20 μg/μL), and the cDNA was digested with a SfiI restriction enzyme and size fractioned following the manufacturer's instructions (CLONTECH Laboratories).

Table 1
List of taxa used in phylogenetic analyses and the number of unique contigs generated from each cDNA library sequenced. Abbreviated names are used in some tables for brevity, but all figures use the names listed in the right most column to demonstrate the higher level relationships.

Species	No. of clones sampled	No. of unique contigs	Abbr. name	Family	Superfamily	Taxon name used in phylogenies
Neodiprion sertifer	2000	795	Ns	Diprionidae	Tenthredinoidea	Symphyta
Campoletis sonorensis	2000	761	Cs	Ichneumonidae	Ichneumonoidea	Ichneumonidae
Lysiphlebus testacipes	n/a	n/a	Lt	Braconidae	Ichneumonoidea	Braconidae
Pristaulacus strangliae	2000	581	Ps	Aulacidae	Evanioidea	Evanioidea
Pelecinus polyturator	3000	842	Pp	Pelecinidae	Proctotrupoidea	Proctotrupoidea
Eucoiliinae sp.	2500	536	Fe	Figitidae	Cynipoidea	Cynipoidea
Nasonia vitripennis	n/a	n/a	Nv	Pteromalidae	Chalcidoidea	Chalcidoidea
Ceraphronidae sp.	2500	492	Ce	Ceraphronidae	Ceraphronoidea	Ceraphronoidea
Apis mellifera	n/a	n/a	Am	Apidae	Apoidea	Apoidea
Solenopsis invicta	n/a	n/a	Si	Formicidae	Vespoidea	Vespoidea
Tribolium casteneum	n/a	n/a	Tc	Tenebrionidae	Tenebrionoidea	Coleoptera
Bombyx mori	n/a	n/a	Bm	Bombycidae	Bombycoidea	Lepidoptera
Drosophila melanogaster	n/a	n/a	Dm	Drosophilidae	Ephydroidea	Diptera
Acyrthosiphon pisum	n/a	n/a	Ap	Aphididae	Aphidoidea	Hemiptera
Myzus persicae	n/a	n/a	Am	Aphididae	Aphidoidea	Hemiptera
Locusta migratoria	n/a	n/a	Lm	Acrididae	Acridoidea	Orthoptera

The cDNA libraries were ligated to λ TriplEx2™ vector in a packaging reaction using PhageMaker® System (Novagen), following the manufacturer's instructions. Phage transductions were performed for 2 h at 31 °C using the BM25.8 *E. coli* host strain in LB broth with 10 mM $MgSO_4$. The converted library was then plated on LB agar plates containing carbenicillin (50 µg/ml) and grown overnight at 37 °C. Isolated colonies were sampled and placed into 96-well PCR plates containing 50 µL of LB broth with 8% glycerol and carbenicillin (50 µg/ml) and grown overnight at 37 °C. The individual colonies were then sampled and picked into 20 µL of water and heated at 95 °C for 2 min. This mixture (2 µL) was then used as template in a 25 µL PCR reaction with 2 nM of TripleX 5LD (5′-CTC GGG AAG CGC GCC ATT GTG TTG GT-3′), 2 nM of Triplex 3LD (5′-TAA TAC GAC TCA CTA TAG GGC GAA TT-3′), 1.25 mM dNTP ~40 U of in-house developed Taq (for method, see Pluthero, 1993), 10× PCR buffer (500 mM KCl, 100 mM Tris–HCl (pH 9.0), and 1% Triton-X-100), and 1.2 mM $MgCl_2$. Thermocycler conditions were as follows: 3 min at 94 °C followed by 32 cycles of 30 s at 94 °C, 30 s at 60 °C, and 1 min at 72 °C, with a final extension of 7 min at 72 °C. Amplified samples were electrophoresed in 1% agarose gel alongside a 1 kb ladder and all reactions demonstrating single bands above 200 bp were sent to the Advanced Genetic Technologies Center, University of Kentucky, for sequencing. Product purification was performed using Agencourt CleanSEQ magnetic beads, and sequencing was carried out using BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) with reaction products analyzed on an Applied Biosystems 3730xl DNA Analyzer.

3.3. Contig assembly and identification of orthologs

All sequences were cleaned of vector contaminants and trimmed of low quality sequence using the program SeqMan (DNASTAR Inc., Madison, WI, USA). These cleaned reads were submitted to dbEST, NCBI. Accession numbers are listed in Section 7. Subsequently, single pass reads were assembled into non-redundant contigs in SeqMan using the default high stringency settings. In addition to the six species of Hymenoptera analyzed here, the predicted genes of three annotated model genomes were utilized, including: *Drosophila melanogaster* (Diptera), *Bombyx mori* (Lepidoptera), and *Apis mellifera* (Hymenoptera). These coding sequences were downloaded from the following resources: Flybase (The FlyBase Consortium, 2008; Tweedie et al., 2009), SilkDB (Beijing Genomics Institute, 2006; Wang et al., 2005), and BeeBase (Elsik et al., 2006; The Honeybee Genome Sequencing Consortium, 2008), respectively.

As an initial search to identify orthologs, we utilized a predeveloped semi-automated software program designed for identifying orthologs of proteomes (Robbertse et al., 2006). All sequences were translated into amino acid sequences and run through the pipeline (Robbertse et al., 2006) which included: a Basic Local Alignment Search Tool (BLASTp) (Altschul et al., 1990) comparing all sequences to all sequences ("all-versus-all") with a cutoff e-value of 1e-1; clustering with MCL (http://micans.org/mcl/) across several inflation parameters (Enright et al., 2002); and cluster filtering. Filtering involved selecting clusters containing proteins that had best hits to other proteins within that same cluster. Additionally, clusters were excluded if it contained more than one protein per species. At minimum, 5 of the 9 taxa had to be included in each cluster. A total of 76 clusters were identified using the pipeline.

When there are hundreds of proteins for each taxon, this high-throughput method of identifying orthologs is extremely efficient. If paralogous sequences seep into the dataset, the conflicting phylogenetic signal is likely to be swamped out by the hundreds of orthologous genes. However, when there are fewer sequences for each taxon, paralogy can contribute significant noise to the dataset and potentially affect the outcome. Thus, to further prevent out-paralogs, the sequences from each cluster were filtered through another set of criteria. Each nucleotide sequence from each cluster was subject to a tBLASTx search against the Reference mRNA sequences (refseq_rna, NCBI) with a higher cutoff e-value of 1e-25 (Altschul et al., 1990). We chose this higher value based on the work of Savard et al. (2006) in hopes that the higher cutoff would minimize spurious sequence similarities. In theory, orthologs should score higher with each other than any other sequence in a given genome (Tatusov et al., 1997). Thus, similar to criteria outlined by Tatusov et al. (1997), each sequence had to have the reciprocal best hit (RBH) for three different model genomes: *D. melanogaster*, *B. mori*, and *A. mellifera*. To prevent the inclusion of short, domain-level matches, the best hits had to have an identity of greater than 50% over a minimum of 60 amino acids. This criteria is similar to the overlap cutoff described in Remm et al. (2001). Additionally, genes were excluded if multiple genes hit below an e-value of 1e-25 for any of these taxa. This effectively excluded large gene families with long conserved domains such as histones. These additional criteria reduced the list of putative orthologs from 76 to 29.

Additional sequences were assigned to the cluster from the following seven taxa if they also met the above criteria: *Nasonia vitripennis* (Hymenoptera), *Solenopsis invicta* (Hymenoptera: Formicidae), *Lysiphlebus testacipes* (Hymenoptera: Braconidae) *Tribolium castaneum* (Coleoptera), *Myzus persicae* (Hemiptera: Aphididae),

Acyrthosiphon pisum (Hemiptera: Aphididae), and *Locusta migratoria* (Orthoptera). These sequences came from the following databases: refseq_rna, non-redundant nucleotide collection (nr/nt, NCBI), and the Non-human, non-mouse ESTs (est_others, NCBI). These taxa increased sampling within the ingroup and provided multiple outgroups for the analysis. To minimize the amount of missing data, clusters were included only if they contained representative sequences from at least three of the six hymenopteran taxa sequenced for this experiment. While 29 of the 76 clusters met the stricter search criterion, only 12 of these contained at least three of the sequenced hymenopteran taxa.

Since the pipeline retains clusters with only one sequence per species, potentially useful genes are eliminated as some taxa possess multiple transcript variants or in-paralogs. Transcript variants often do not vary across the coding sequence or differ only in one or a few sites that will likely will not affect the overall phylogenetic analysis (Goodstadt and Ponting, 2006). Additionally, in-paralogs are lineage specific gene duplications that are orthologs by definition (Remm et al., 2001). Similar to transcript variants, in-paralogs should be more similar within species than between (Kuzniar et al., 2008; Remm et al., 2001), and the assumption is that their inclusion will not affect the phylogenetic analysis. However, out-paralogs are gene duplications that occurred before a given speciation event, and the copies typically take on a different function than the original (Kuzniar et al., 2008). Thus, the sequences of out-paralogs should be well-differentiated from the original copy, and thus can be identified with phylogenetic analysis.

To increase the number of genes available for analysis, all sequences from the six hymenopteran libraries were again examined using an all versus all blastn search (Altschul et al., 1990) with a cutoff e-value of 1e-25 using the stand alone blastall program (NCBI). All hits that were not identified with the pipeline (due to multiple sequences per specimen) were filtered using the same criteria and methodology mentioned previously. An additional 12 genes were identified, all with at least one taxon having multiple transcript variants or putative in-paralogs. All sequences were aligned using MUSCLE (Edgar, 2004), and hand edited to ensure a proper reading frame. To test whether genes with multiple transcripts were useful and did not represent out-paralogs, all transcripts for all taxa were tested phylogentically (see description in Section 3.4 below). If the transcripts for a given taxon clumped together on the tree, they were considered transcript variants or in-paralogs and therefore, were included within the dataset provided they met all other criteria. The final dataset consisted of 24 genes, 12 identified from the pipeline and 12 identified through the method just described.

3.4. Phylogenetic Inference

The number of informative sites and tests for base composition homogeneity were performed in Paup* 4.0b10 (Swofford, 2000). Phylogenetic assessments of taxa with multiple transcript variants were performed using maximum composite likelihood distances (Tamura et al., 2004) and the neighbor-joining method with MEGA 4.0.2 (Tamura et al., 2007). All analyses performed in Paup* 4.0b10 (Swofford, 2000) were aided with the PaupUp graphical interface (Calendini and Martin, 2005). MrModeltest v2.3 (Nylander, 2004; Posada and Crandall, 1998) was used with Paup* 4.0b10 (Swofford, 2000) and the ModelTest Server (Posada, 2006) to test for the best evolutionary model applicable to all datasets and partitions using the Bayesian information criterion. Concatenated datasets were partitioned by codon position. The general time reversible model had the highest likelihood with a parameter for invariant sites and among-site rate variation modeled with a gamma distribution (GTR+I+G) for all partitions. Bayesian inference was used to analyze all concatenated and individual gene datasets with

MrBayes v3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). All analyses were run with 4 chains and 2 independent runs until stationarity was reached. Stationarity of the independent runs was determined using convergence diagnostics and plots of generation versus the log probability of the data as guidelines. A gene jackknife was performed by creating 100 pseudo-replicates containing 15 genes sampled at random without replacement (62.5% removal probability) using a perl script. Each new dataset was analyzed under the same Bayesian framework described above and the resulting bipartitions from each pseudo-replicate were summarized.

The maximum likelihood analysis was performed on the concatenated dataset with all data included, using RAxML VI-HPC (Stamatakis, 2006) on the CIPRES Portal v. 1.14 (CIPRES Collaborative Group, 2005–2008) with GTRGAMMAI model with rapid bootstrapping (under GTRCAT model) and automatic determination of the number of replications required (Stamatakis et al., 2008). Parsimony analyses were also performed on the full concatenated dataset using Paup* 4.0b10 (Swofford, 2000) with a heuristic search, 1000 random additions sequences, TBR, holding 5 trees per rep, and multiple states treated as polymorphisms. Standard bootstrap resampling was performed with the same heuristic search settings with 1000 replications. Evolutionary networks were constructed using SplitsTree v.4.0 (Huson and Bryant, 2006), with filtered supernetworks performed using the Z-closure method (see Huson et al., 2004 for a detailed explanation). Phylogenetic trees were viewed and manipulated using Dendroscope (Huson et al., 2007).

4. Results

4.1. Concatenated datasets

The final concatenated dataset contained 24 genes with an aligned length of 10,917 base pairs of which 48.6 percent were parsimony informative. Table 2 lists which genes were included in the dataset and which taxa were represented in the individual gene datasets. All of the individual gene datasets had a minimum of 12 taxa with a representative transcript. Under a Bayesian framework, the 24-gene dataset recovered several expected relationships consistent with other molecular and morphological phylogenetic studies of Hymenoptera, including: a monophyletic Apocrita, Aculeata, Ichneumonoidea, and a sister relationship between the two most closely related putative proctotrupomorphs (Cynipoidea + Proctotrupoidea) (Fig. 2). Additionally, the phylogenetic positions of all outgroups were consistent with previously recovered relationships (Savard et al., 2006; Wheeler et al., 2001; Whiting, 2002; Wiegmann et al., 2009), including a monophyletic Holometabola, Hymenoptera as sister to all other Holometabola, and a sister relationship between the two included Panorpoid orders.

Within the ingroup, Ceraphronoidea and Evanioidea were recovered as sister taxa, consistent with Rasnitsyn's (1988) proposed Evaniomorpha. However, this Evaniomorpha clade was recovered in a polytomy with all other apocritan taxa (Fig. 2), rather than as sister to Proctotrupomorpha as proposed by Rasnitsyn (1988) (cf. Fig. 1A). The Proctotrupomorpha was proposed by Rasnitsyn (1988) to include Cynipoidea, Proctotrupoidea *s.l*, Chalcidoidea, and Platygastroidea. Although the platygastroids were not represented in this analysis, Chalcidoidea was not recovered with the other putative proctotrupomorphs. Rather, Proctotrupoidea + Cynipoidea were recovered as sister to Aculeata and Chalcidoidea was recovered in the basal apocritan polytomy.

Table 3 lists which genes recovered the clades depicted in Fig. 2. The node numbers in Table 3 correspond to the circled node labels

Table 2
List of genes used in analyses, including the taxa represented for each gene, the aligned length, the number of parsimony informative (P.I.) sites, and the percent missing data for each taxon. Gene numbers and symbols are referenced to FlyBase (The FlyBase Consortium, 2008). See Table 1 for the key to abbreviated taxon names.

| Flybase gene number	Flybase gene symbol	Aligned length	No. P.I. sites	Outgroups						Hymenopteran taxa									
				Lm	Bm	Dm	Tc	Ap	Mp	Nv	Am	Si	Lt	Cs	Ce	Ns	Fe	Pp	Ps
CG1746	CG1746	444	204	√	√	√	√	√	√	√	√	√	√	√	√	√	–	–	–
CG2099	RpL35A	342	177	√	√	√	√	√	√	√	√	√	√	√	–	√	√	√	–
CG2746	RpL19	612	260	√	√	√	√	√	√	√	√	√	√	√	–	–	√	√	–
CG3186	eIF-5A	486	195	√	√	√	√	√	√	√	√	√	√	√	–	√	√	–	–
CG3446	CG3446	432	339	√	√	√	√	√	√	√	√	√	√	√	–	√	–	√	–
CG3661	RpL23	423	148	√	√	√	√	√	√	√	√	√	√	√	–	√	√	–	√
CG3997	RpL39	156	60	√	√	√	√	√	√	√	√	√	√	√	√	–	√	√	–
CG4097	Pros26	471	262	√	√	√	√	√	√	√	√	√	–	√	√	–	–	–	√
CG4169	CG4169	771	540	√	√	√	√	√	√	√	√	√	√	√	–	√	–	√	–
CG4800	Tctp	531	270	√	√	√	√	√	√	√	√	√	√	√	–	√	√	–	√
CG6770	CG6770	195	101	√	√	√	√	√	√	√	√	√	√	√	√	√	–	√	–
CG6779	RpS3	708	343	√	√	√	√	√	√	√	√	√	√	√	√	–	√	√	√
CG6803	Mf	318	180	√	√	√	√	√	√	√	√	√	√	√	–	√	–	√	√
CG7178	wupA	597	229	√	√	√	√	√	√	√	√	–	√	√	–	–	–	√	√
CG7424	RpL36A	309	119	√	√	√	√	√	√	√	√	√	√	√	–	√	–	√	√
CG7434	RpL22	378	197	√	√	√	√	√	√	√	√	√	√	√	√	–	√	–	–
CG7939	RpL32	405	186	√	√	√	√	√	√	√	√	√	√	√	–	–	√	√	–
CG8332	RpS15	456	192	√	√	√	√	√	√	√	√	√	√	–	–	√	√	√	–
CG8415	RpS23	429	157	√	√	√	√	√	√	√	√	√	√	–	√	√	–	√	√
CG8857	RpS11	471	206	√	√	√	√	√	√	√	√	√	√	√	–	–	√	–	√
CG8900	RpS18	498	186	√	√	√	√	√	√	√	√	√	√	–	–	√	√	–	√
CG11271	RpS12	429	230	√	√	√	√	√	√	√	√	√	√	√	–	–	–	√	√
CG11981	Prosβ3	618	327	√	√	√	√	√	√	√	√	√	–	–	√	√	–	–	√
CG15442	RpL27A	438	198	√	√	√	√	√	√	√	√	√	√	√	√	–	√	√	√
Total No. genes for each taxon				24	24	24	24	24	24	24	24	24	21	20	11	13	14	15	12
Percent missing data				5.8	1.5	1.6	1.3	1.9	4.0	1.1	1.0	10.7	20.9	22.3	66.2	51.0	52.1	49.2	51.7
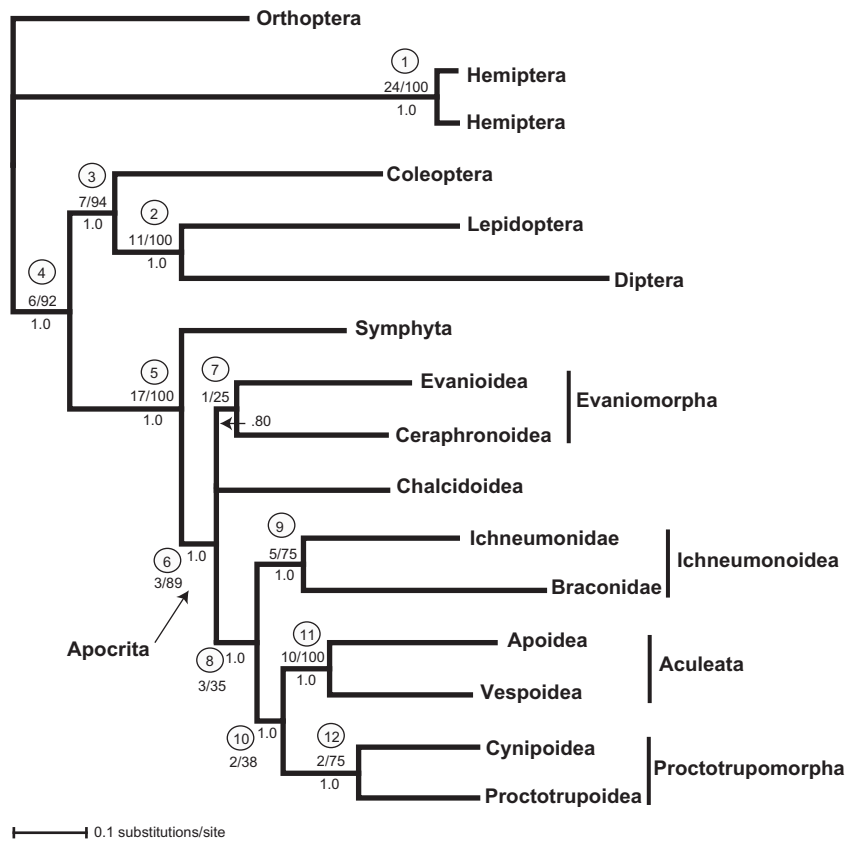
Fig. 2. Bayesian phylogram inferred from the concatenated dataset of 24 genes. (1 million generations, burnin = 150 K generations). The circled numbers above or to the left of a node represent labels for ease of discussion and can be crossed reference with the node labels in Table 3. Posterior probabilities are listed below the node. The number of genes that recovered a clade is listed before the forward slash. The percentage of pseudo-replicates recovering a clade from the gene jackknife analysis is indicated after the forward slash.

Table 3
List of which genes (from individual gene analyses, see Fig. S1, Supplementary information) supported the clades recovered in Fig. 2. Refer to Fig. 2 for clades for the node numbers. A checkmark indicates that node was recovered in the individual gene analysis, whereas a blank cell indicates the node was not recovered. A gray cell indicates that node could not be recovered due to missing taxa. (A) Total number of genes supporting clade; (B) Total number of genes possible for clade recovery; (C) Percent of genes supporting clade.

Gene	1	2	3	4	5	6	7	8	9	10	11	12
CG1746	√						—		√	—	√	—
CG2099	√	√			√		—				√	
CG2746	√	√	√		√	—						√
CG3186	√		√	√	√		—					—
CG3446	√	√	√				—					—
CG3661	√	√					—					—
CG3997	√											
CG4097	√				√	—			—		√	—
CG4169	√				√		—					—
CG4800	√								√		√	—
CG6770	√				√		—					—
CG6779	√	√		√	√	—						
CG6803	√	√	√	√		√	—		√		√	
CG7178	√				√	—	—	√	—	√	√	—
CG7424	√				√	—	—		√			
CG7434	√						—		√			—
CG7939	√		√		√	—	—					
CG8332	√		√	√	√	√	—		√	—	√	
CG8415	√				√	√			√	—		—
CG8857	√	√	√	√	√	—	—				√	–
CG8900	√	√					—		—			—
CG11271	√	√			√	—	—				√	—
CG11981	√	√			√		√		—	—	—	—
CG15442	√	√		√	√	—				√	√	√
A	24	11	7	6	17	3	1	3	5	2	10	2
B	24	24	24	24	24	14	7	23	18	22	24	8
C	100	46	29	25	71	21	14	13	28	9.1	42	25

depicted in Fig. 2. Fig. 2 depicts the number of genes that recovered a given node (before the slash), determined from examining the recovered clades from individual gene analyses (Fig. S1, Supplementary information). The number after the slash is the percentage of pseudo-replicates that recovered that node in the gene jackknife analysis. Although there was high support (posterior probability (pp) >0.95) over most of the tree, there was relatively low nodal support for Evaniomorpha (node 7, pp = 0.80). Evaniomorpha was recovered in only one gene tree and in only 25 percent of the gene jackknife pseudo-replicates. Interestingly, Ceraphronoidea was recovered more commonly with Cynipoidea + Proctotrupoidea (36%, data not shown) than with Evanioidea (25%) across the gene jackknife pseudo-replicates. This demonstrates the limited evidence placing Ceraphronoidea as sister to Evanioidea. It is possible that sampling error affected node recovery and support, particularly for the Evaniomorpha clade, as both the ceraphronoid and evanioid had 66.2 and 51.7 percent missing data (including gaps), respectively (Table 2).

The highest supported node in terms of the percent of genes possible for clade recovery was unsurprisingly between the two most closely related taxa, the two hemipterans (node 1). The next highest supported node was the hymenopteran clade (node 5), with 17 out of the 24 genes indicating monophyly. While this clade

has never been in doubt morphologically, the numerous genes recovering this and other clades (e.g. Diptera + Lepidoptera) reveal the phylogenetic potential of these loci for higher level phylogenetics of insects. Node 10, which represents a sister relationship between Aculeata and Cynipoidea + Proctotrupoidea, had a high posterior probability (1.0) but had the lowest percent of possible genes supporting the clade (2 out of 22, Table 3), and was only recovered in 38 of the gene jackknife pseudo-replicates. Similarly, the sister relationship between Ichneumonoidea and Aculeata + Proctotrupomorpha (node 8) was only recovered in 3 gene trees and in 35 gene jackknife pseudo-replicates (Fig. 2). Neither of these clades has been recovered by previous molecular, morphological, or combined analyses (Castro and Dowton, 2006; Dowton and Austin, 1994, 2001; Ronquist et al., 1999). However, it should be noted that Rasnitsyn (1980) had previously proposed a close relationship between Proctotrupomorpha, Aculeata, and Ichneumonoidea (as recovered, node 8) based on the shared presence of articulating propodeal condyles.

Given the disparity in branch lengths among the outgroup taxa, it is possible that outgroup rooting affected the result. To test for the effect of outgroup selection, three different analyses were performed. First, the orthopteran was excluded and the analysis was rooted on A. pisum (Hemiptera). Second, all outgroups were ex-
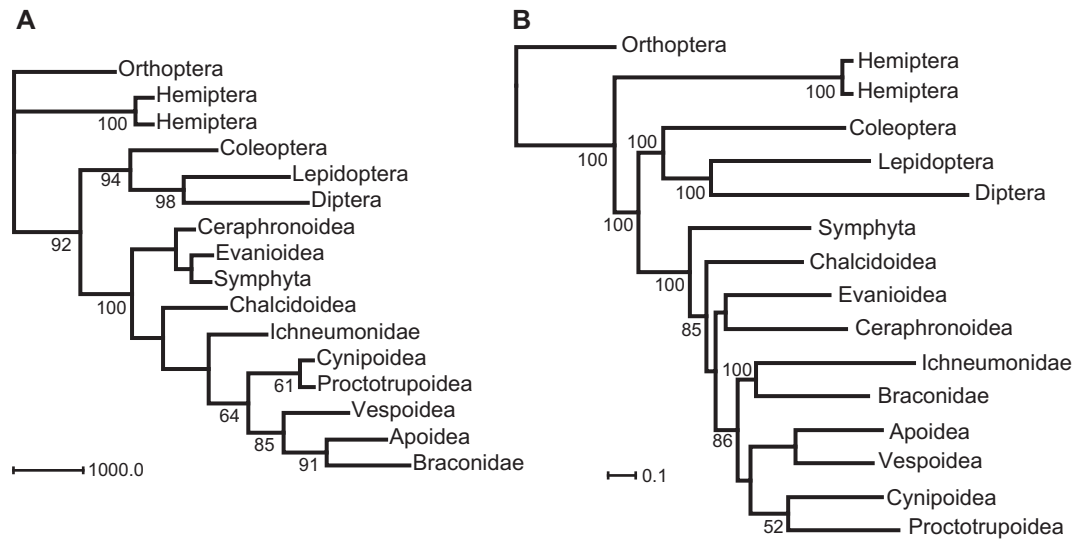
Fig. 3. (A–B). Parsimony and maximum likelihood analyses of the concatenated 24-gene dataset. Bootstrap values (>50) are listed below the node. (A) Single most parsimonious tree (Length = 21,655, Consistency index = 0.48, Retention index = 0.38). (B) Phylogeny with the highest likelihood (−89026.338715) with model GTRGAMMAI, and partitioned by codon position. The scale bar represents number of substitutions per site.

cluded except for the orthopteran, thereby excluding potential effects from the long branches of the panorpoid orders and the hemipterans. Finally, all outgroups were excluded except the coleopteran, potentially reducing the divergence time between the ingroup and outgroup. Regardless of outgroup selection and inclusion, all three analyses produced the same topology in Fig. 2.

Parsimony and maximum likelihood analyses were performed to test if phylogenetic method affected the results. A parsimony analysis of the concatenated dataset recovered one most parsimonious tree (Fig. 3A). Additionally, a maximum likelihood analysis was performed and the resulting phylogeny is depicted in Fig. 3B. The maximum likelihood analysis produced a similar tree to the Bayesian analysis depicted in Fig. 2, although Chalcidoidea was recovered as sister to the remaining apocritans rather than in a basal apocritan polytomy. All outgroup relationships were the same across all inference methods and Hymenoptera was monophyletic. While the likelihood analysis recovered a monophyletic Apocrita, Aculeata, and Ichneumonoidea, the parsimony analysis did not. None of the inference methods placed the Chalcidoidea in a clade with Cynipoidea and Proctotrupoidea, contrary to Rasnitsyn's (1988) concept of Proctotrupomorpha. Regardless of method, Ichneumonoidea, Aculeata, and Cynipoidea + Proctotrupoidea were recovered together in a clade, although the branching order was altered in the parsimony analysis. Given the extreme A-T bias for both Apoidea and Braconidae in the third position relative to the other taxa (Table S1, Supplementary information), it is most likely that the braconid was misplaced in the parsimony analysis (Fig. 3A), causing the unexpected paraphyly of Aculeata and Ichneumonoidea.

4.2. Nucleotide composition bias

To test if a nucleotide composition bias affected the analysis, chi-square tests for base composition homogeneity were performed (Table S2, Supplementary information). For the individual gene alignments, 22 out of 24 genes failed the test for base composition homogeneity ($p < 0.05$). Interestingly, when the dipteran was excluded from the test (which possessed the longest branch lengths across most topologies), only 10 of the 24 genes failed the homogeneity test with all data included (data not shown).

The concatenated dataset also demonstrated a lack of base composition stationarity (Table S2, Supplementary information). Each

gene and the concatenated dataset were tested for nucleotide composition homogeneity for each codon position and with only the third position excluded. The null hypothesis of homogeneity was accepted for all genes with the third position excluded ($p < 0.05$), but not for the concatenated dataset (Table S2, Supplementary information), indicating potential systematic error. Only one gene (CG4169) failed the test for the first codon position (Table S2, Supplementary information) and homogeneity was indicated for all genes and for the fully aligned dataset for the second position (data not shown).

To reduce any potential effects from nucleotide composition biases on the phylogenetic inference, all positions indicating heterogeneity for each individual gene were excluded in a concatenated analysis (Mix-P). The Mix-P dataset had the 3rd position excluded for 21 of the genes indicating compositional heterogeneity for this codon position, and the gene indicating heterogeneity in the 1st and 3rd positions (CG4169) was translated to amino acids and separated into its own partition under a mixed model in MrBayes. Bayesian inference of this dataset recovered the exact same topology as the maximum likelihood analysis depicted in Fig. 3B. All nodes had a posterior probability of 1.0 except for Evaniomorpha (pp = 0.95) and the clade containing Evaniomorpha and the remaining apocritans (pp = 0.53). The low support for this latter node highlights the uncertainty in the position of Evaniomorpha relative to Chalcidoidea in the apocritan lineage.

4.3. Individual gene analyses

Individual gene trees displayed a lack of concordance with the concatenated analysis and with each other (Fig. S1, Supplementary information). Several gene trees had very little resolution or very low nodal support for internal branches. This is not surprising given the short length and conserved nature of each gene and the low taxonomic sampling. Although 3 of the gene trees (CG6803, CG7178, and CG7939) were compatible with the ingroup relationships recovered in the concatenated analysis depicted in Fig. 2, two of these trees had very little resolution.

Given that almost all of the genes violated the assumption of base composition homogeneity in the third position, each gene was reanalyzed with the third position removed if it failed the homogeneity test (Fig. S2, Supplementary information). Additionally, gene CG4169 was analyzed as a protein since the first codon

position also failed the homogeneity test. This mixed inclusion of sites across the different genes and the analysis of gene CG4169 as a protein was the data included within the Mix-P dataset, discussed earlier in Section 4.2. Comparing the individual gene trees in Fig. S2 (Supplementary information) to Fig. 2, only 2 of the gene trees were compatible with the ingroup relationships. One of these gene trees (CG3661) did not have any resolution in the ingroup (Fig. S2, Supplementary information). The other gene tree (CG15442) was only resolved for Cynipoidea + Proctotrupoidea within the ingroup (Fig. S2, Supplementary information). However, all highly supported nodes were compatible with the clades recovered in Fig. 2. This was not the case when all data was included, as several gene trees recovered highly supported nodes that conflicted with the topology in Fig. 2.

Interestingly, there were 5 genes that indicated a sister relationship between Apoidea and Braconidae and 5 genes that recovered the accepted sister relationship between Ichneumonidae and Braconidae when all data were included (Fig. S1, Supplementary information). When the third position was removed from genes with heterogeneous base composition, not one of the genes recovered the erroneous Braconidae + Apoidea relationship (Fig. S2, Supplementary information). Both of these taxa had similar A-T composition across a number of genes, and their recovery together in some individual gene trees was likely due to the convergent evolution of these nucleotides at the third position.

4.4. Visualizing conflict and compatibility with filtered supernetworks

Filtered supernetworks (Huson and Bryant, 2006) have been successfully used to visualize the most common relationships given a set of taxonomically overlapping gene trees (Whitfield et al., 2008), an especially useful tool when there is a high degree of conflict among the input trees. Fig. 4 illustrates filtered supernetworks, which include only those splits contained in (or compatible with) a set minimum number of trees. Fig. 4A–C depicts supernetworks created from the 24 individual gene topologies. As shown in Fig. 4A, the split representing Aculeata (i.e. Apoidea + Vespoidea) was contained in more than 50% of the gene trees (Min. trees = 13). Most apocritan lineages were separated from Symphyta, but the position of Ceraphronoidea was reticulated with respect to outgroups and the ingroup. Clearly, Ceraphronoidea demonstrated the most conflicting phylogenetic positions across the individual gene datasets.

Only two gene trees recovered Aculeata as sister to Proctotrupomorpha (i.e. Cynipoidea + Proctotrupoidea) (Fig. 2). However, this clade is recovered in the supernetwork when the filter is set to 11 minimum trees (Fig. 4B, see arrow), demonstrating that there is more evidence for this relationship than can be readily determined from examining individual gene trees. Across the individual gene trees, the relationship between Aculeata and Proctotrupomorpha was likely obscured in part by the similar A-T composition bias possessed by Apoidea and Braconidae. As mentioned previously in Section 4.3, there were several individual gene trees that erroneously placed Apoidea as sister to Braconidae, due to suspected convergent nucleotide composition. When the braconid along with the volatile ceraphronid is filtered out of the supernetwork, the split containing Aculeata + Proctotrupomorpha is recovered in a 14 minimum tree filtered supernetwork (see arrow, Fig. 4C).

The 100 pseudo-replicates from the gene jackknife were also analyzed using filtered supernetworks (Fig. 4D–F). When the filter was set to 50 trees minimum, the splits representing Apocrita, Ichneumonoidea, Aculeata, and Proctotrupomorpha are all recovered, as would be expected given the high ($\geqslant 75$) gene jackknife support for these clades (Fig. 2). However, when the volatile Ceraphronoidea is filtered out, the split containing (Ichneumonoidea + Aculeata + Proctotrupomorpha) is recovered within 50 trees (Fig. 4E, see arrow); even though these additional clades have low gene jackknife support (<40) (cf. node 9 and node 11, Fig. 2). Thus, the volatile placement of Ceraphronoidea lowers the gene jackknife support for these relationships. When the braconid is also filtered out, the split containing Aculeata + Proctotrupomorpha is recovered in 70 minimum trees (Fig. 4F).

5. Discussion

The phylogenetic potential of these loci was revealed by the consistent recovery of all well corroborated evolutionary relationships. Under maximum likelihood and Bayesian inference, all analyses of the concatenated nucleotides recovered a monophyletic Holometabola, Hymenoptera, Apocrita, Aculeata, Ichneumonoidea, and a sister relationship between the two most closely related proctotrupomorphs (Cynipoidea + Proctotrupoidea). Recovery of these relationships was robust to missing data, nucleotide composition biases, conflicting signal across gene trees, and low taxonomic sampling. Thus, ESTs have great potential for resolving higher-level Hymenopteran relationships as well as ordinal relationships among insects, which will likely become more apparent with greater taxonomic and genetic sampling.

The volatile placement of many taxa was evident across the individual gene trees. Ceraphronoidea displayed the greatest level of reticulation across splits with respect to both the ingroup and outgroups (Fig. 4A–B). Braconidae also acted as a rogue taxon in individual gene trees, often pairing with Apidae due to convergent A-T nucleotide composition. As visualized by the gene jackknife and gene tree filtered supernetworks (Fig. 4A–F), volatile taxa clearly obscured some of the historical signal, a problem likely over-exaggerated with low taxonomic sampling. However, the relative placement of Chalcidoidea with respect to Evanioidea and Ceraphronoidea could not be readily determined with this dataset.

Since there was only one exemplar for each major lineage, it is possible that the low taxonomic sampling played a role in the lack of congruence among individual gene datasets. For a given gene, the pattern and rate of substitution of a given taxon may not have been characteristic for the group it represents. Thus, some gene trees may have recovered historical relationships, some may have had insufficient signal, and others may have recovered false relationships due to long-branch attraction artifacts and biases in the pattern of substitution (Collins et al., 2005). Increased taxonomic sampling has been the most common and effective remedy for both phylogenetic conflict (Dunn et al., 2008; Hedtke et al., 2006) and long-branch attraction (Bergsten, 2005), and is the obvious next step for future empirical studies using ESTs for hymenopteran relationships.

5.1. Phylogenetic implications

One of the most interesting results of this study is the sister relationship between Proctotrupomorpha and Aculeata. Previous studies have either recovered Aculeata within Evaniomorpha (Castro and Dowton, 2006) or more commonly, as sister to Ichneumonoidea (Brothers, 1975; Dowton and Austin, 2001; Oeser, 1961; Rasnitsyn, 1988; Vilhelmsen et al., 2010). However, the support for these relationships in previous studies has been tenuous. Here, we propose Proctotrupomorpha (not including Chalcidoidea) as the sister group to Aculeata. This relationship was stable and highly supported with the exclusion and inclusion of various outgroups and data. Only two genes recovered this relationship with all data included (figure S1, Supplementary information), however, there were several more gene trees compatible with this relationship, as shown in the filtered supernetworks (Fig. 4).
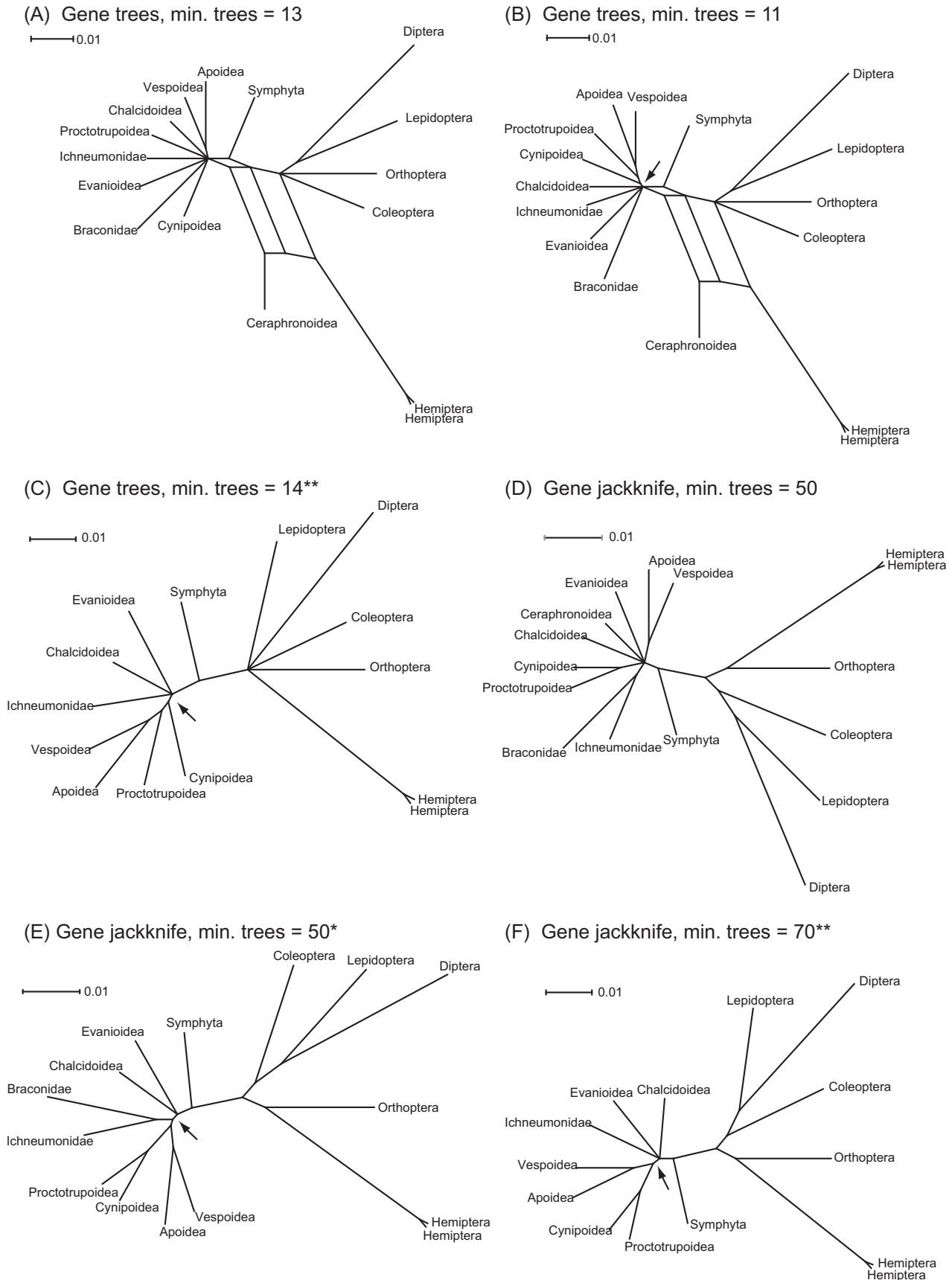
Fig. 4. (A–C) Filtered supernetworks of the 24 individual gene trees. (D–F) Filtered supernetworks of the 100 gene jackknife pseudo-replicates. Each supernetwork includes only those splits contained in (or compatible with) a set minimum number of gene trees, as indicated in the figure subtitle. *Ceraphronoidea excluded. **Braconidae and Ceraphronoidea excluded. Arrows demonstrate splits inferring relationships between Aculeata (i.e. Apoidea + Vespoidea) and Proctotrupomorpha (i.e. Proctotrupoidea + Cynipoidea) and Ichneumonoidea (i.e. Ichneumonidae + Braconidae). The scale bar represents number of substitutions per site.

There is strong evidence suggesting that Chalcidoidea does not belong within the Proctotrupomorpha, as suggested by Gibson (1999) and more recently supported with additional morphological evidence (Vilhelmsen et al., 2010). However, much greater taxon sampling will be required to fully understand the placement of Chalcidoidea and its closest relatives, given the lack of consistent placement of this taxon across the different analyses. The support for the placement of Chalcidoidea as sister to all remaining apocritans was poorly supported in both the maximum likelihood analysis and the Mix-P dataset (Fig. 3B).

The placement of Evanioidea and Ceraphronoidea also cannot be determined with certainty on the available evidence. Although they were recovered as sister taxa in most concatenated analyses, the relationship was poorly supported and their relative positions were more sensitive to method of inference. Interestingly, across a vast majority of the individual gene trees, Chalcidoidea, Ceraphronoidea, and Evanioidea, demonstrated an earlier divergence from the remaining apocritans (Figs. S1–S2, Supplementary information). The high levels of missing data in the latter two taxa likely contributed to their highly volatile placements across the different gene trees. Thus, resolving the true phylogenetic positions of these taxa will require further genetic and taxonomic sampling.

6. Conclusions

From this study, it is evident that ESTs have great potential to resolve higher-level hymenopteran relationships. Even though holometabolan relationships were not the focus of this study, given the accurate resolution across the included orders, it is also clear that ESTs will be very useful for resolving long contested ordinal relationships. ESTs allow for greater taxonomic sampling beyond model organisms from genome projects and a more comprehensive phylogenomic study of Hymenoptera would be a prudent future study. Additionally, future studies should take advantage of next-generation sequencing technologies that will make large scale sampling a realistic and cost-effective pursuit. One recommendation would be to normalize the cDNA libraries in the future for a more robust sampling in the RNA pool. Increased taxonomic sampling can also be achieved with primer design from the alignments produced from this study to amplify nuclear DNA of rare and previously collected specimens. Studies are currently underway for utilizing novel genes for understanding macroevolution across Hymenoptera (Sharanowski, unpublished data) and have been completed for population genetic and phylogeographic studies in gall wasp parasitoids (Lohse et al., 2010). These loci demonstrate great promise for finally understanding the evolution of this extremely diverse and important group of insects.

The data and analyses performed herein points to a sister relationship between Aculeata and Proctotrupomorpha (i.e. Proctotrupoidea + Cynipoidea) contrary to previously proposed hypotheses. Most evidence demonstrates that Ichneumonoidea is also closely related to Aculeata + Proctotrupomorpha. Additionally, there is evidence for the antiquity for both the evanioid and ceraphronoid lineages. Most of the available evidence suggests that Chalcidoidea is not contained within Proctotrupomorpha. These results have important evolutionary implications for morphological and behavioral character studies as well as for other phylogenetic studies for appropriate outgroup selection.

7. dbEST and Dryad accession numbers

Data files are available at the Dryad Digital Repository (http://datadryad.org/) under DOI:10.5061/dryad.1735. Genbank accession numbers for all EST sequences are HO079272–HO087893.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.07.006.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Austin, A., Dowton, M., 2000. Hymenoptera: Evolution, Biodiversity, and Biological Control. CSIRO, Canberra.
Barnes, W.M., 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. PNAS 91, 2216–2220.
Baurain, D., Brinkmann, H., Philippe, H., 2007. Lack of resolution in the animal phylogeny: closely spaced cladogenesis or undetected systematic errors? Mol. Biol. Evol. 24, 6–9.
Beijing Genomics Institute, 2006. Silkworm Knowledge Base: SilkDB. Available from: <http://silkworm.genomics.org.cn/> (accessed June 2008).
Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163–193.
Brothers, D.L., 1975. Phylogeny and classification of the aculeate Hymenoptera, with special reference to Mutillidae. Univ. Kansas Sci. Bull. 50, 483–648.
Calendini, F., Martin, J.-F., 2005. PaupUP v1.0.3.1: a free graphical frontend for Paup* Dos software.
Carpenter, J.M., Wheeler, W.C., 1999. Towards simultaneous analysis of molecular and morphological data in Hymenoptera. Zool. Scr. 28, 251–260.
Castro, L., Dowton, M., 2006. Molecular analyses of the Apocrita (Insecta: Hymenoptera) suggest that the Chalcidoidea are sister to the diapriod complex. Invertebr. Syst. 20, 603–614.
Castro, L., Dowton, M., 2007. Mitochondrial genomes in the Hymenoptera and their utility as phylogenetic markers. Syst. Entomol. 32, 60–69.
Chenchik, A., Zhu, Y.Y., Diatchenko, L., Li, R., Hill, J., et al., 1998. Generation and use of high-quality cDNA from small amounts of total RNA by SMART PCR. In: Siebert, P.D., Larrick, J.W. (Eds.), Gene Cloning and Analysis by RT-PCR. BioTechniques Books, Westborough, pp. 305–319.
Chomczynski, P., Sacchi, N., 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. Anal. Biochem. 162, 156–159.
CIPRES Collaborative Group, 2005–2008. CIPRES: Cyberinfrastructure for Phylogenetic Research. San Diego Supercomputing Center. Available from: <http://www.phylo.org/> (accessed 2008).
Collins, T.M., Fedrigo, O., Naylor, G.J.P., 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogentics. Syst. Biol. 54, 493–500.
Dávalos, L.M., Perkins, S.L., 2008. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91, 433–442.
Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2 (5), e68.
Dowton, M., Austin, A., 1994. Molecular phylogeny of the insect order Hymenoptera: apocritan relationships. PNAS 91, 9911–9915.
Dowton, M., Austin, A., 2001. Simultaneous analysis of 16S, 28S, COI and morphology in the Hymenoptera: apocrita-evolutionary transitions among parasitic wasps. Biol. J. Linn. Soc. 74, 87–111.
Dowton, M., Austin, A.D., Dillon, N., Bartowsky, E., 1997. Molecular phylogeny of the apocritan wasps: the Proctotrupomorpha and Evaniomorpha. Syst. Entomol. 22, 245–255.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Elsik, C.G., Worley, K.C., Zhang, L., Milshina, N.V., Jiang, H., et al., 2006. Community annotation: procedures, protocols, and supporting tools. Genome Res. 16, 1329–1333.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.

Gauld, I., Bolton, B., 1988. The Hymenoptera. British Museum of Natural History, Oxford University Press, London.

Gibson, G.A.P., 1999. Sister-group relationships of the Platygastroidea and Chalcidoidea (Hymenoptera) Ð an alternate hypothesis to Rasnitsyn (1988). Zool. Scr. 28, 125–138.

Goodstadt, L., Ponting, C.P., 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comp. Biol. 2, e133.

Hedtke, S., Townsend, T., Hillis, D., 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55, 522–529.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

Hughes, J., Longhorn, S., Papadopoulou, A., Theodorides, K., de Riva, A., et al., 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (Beetles). Mol. Biol. Evol. 23, 268–278.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Huson, D.H., Dezulian, T., Kloepper, T., Steel, M., 2004. Phylogenetic super-networks from partial trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 1, 151–158.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., et al., 2007. Dendroscope – an interactive viewer for large phylogentic trees. BMC Bioinformatics 8, 460.

Königsmann, E., 1976. Das phylogenetische system der Hymenoptera. Tiel 1: Einführung, Grundplanmerkmale, Schwestergruppe und Fossilfunde. Dtsch. Entomol. Zeitung 23, 253–279.

Königsmann, E., 1978a. Das phylogenetische system der Hymenoptera. Tiel 3: Terebrantes (Unterordnung Apocrita). Dtsch. Entomol. Zeitung 25, 1–55.

Königsmann, E., 1978b. Das phylogenetische system der Hymenoptera. Tiel 4: Aculeata (Unterordnung Apocrita). Deutsche Entomologische Zeitung 25, 365–435.

Kuzniar, A., van Ham, R.C.H.J., Pongor, S., Leunissen, J.A.M., 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 24, 539–551.

Lohse, K., Sharanowski, B.J., Stone, G.N., 2010. Quantifying the Pleistocene history of the oak gall parasitoid Cecidostiba fungosa using twenty intron loci. Evolution, in press, doi:10.1111/j.1558-5646.2010.01024.x.

Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Oeser, R., 1961. Vergleichend-morphologische Untersuchungen über den Ovipositor der Hymenopteren. Mitteilungen Zool. Museum Berl. 37, 1–119.

Philippe, H., Lartillot, N., Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253.

Pluthero, F.G., 1993. Rapid purification of high-activity Taq DNA polymerase. Nucleic Acids Res. 21, 4850–4851.

Posada, D., 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucelotide substitution. Nucleic Acids Res. 34, W700–W703.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Rasnitsyn, A.P., 1980. The origin and evolution of Hymenoptera. Trans. Paleontol. Inst. Acad. Sci. USSR 174, 1–192.

Rasnitsyn, A.P., 1988. An outline of evolution of hymenopterous insects (order Vespida). Orient. Insects 22, 115–145.

Remm, M., Storm, C.E.V., Sonnhammer, E.L.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052.

Robbertse, B., Reeves, J.B., Schoch, C.L., Spatafora, J.W., 2006. A phylogenomic analysis of the Ascomycota. Fungal Genet. Biol. 43, 715–725.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Ronquist, F., Rasnitsyn, A.P., Roy, A., Eriksson, K., Lindgren, M., 1999. Phylogeny of the hymenoptera: a cladistic reanalysis of Rasnitsyn's (1988) data. Zool. Scr. 28, 13–50.

Savard, J., Tautz, D., Richards, S., Weinstock, G.M., Gibbs, R.A., et al., 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16, 1334–1338.

Schulmeister, S., Wheeler, W.C., Carpenter, J.M., 2002. Simultaneous analysis of the basal lineages of Hymenoptera (Insecta) using sensitivity analysis. Cladistics 18, 455–484.

Sharkey, M.J., 2007. Phylogeny and classification of Hymenoptera. Zootaxa 1668, 521–548.

Sharkey, M.J., Roy, A., 2002. Phylogeny of the Hymenoptera: a reanalysis of the Ronquist et al. (1999) reanalysis, emphasizing wing venation and apocritan relationships. Zool. Scr. 31, 57–66.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 75, 758–771.

Swofford, D.L., 2000. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Sunderland, Mass.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101, 11030–11035.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families (cover story). Science 278, 631.

The FlyBase Consortium, 2008. FlyBase: A Database of Drosophila Genes & Genomes. Available from: <http://flybase.org/> (accessed June 2008).

The Honeybee Genome Sequencing Consortium, 2008. BeeBase: Hymenoptera Genome Database. Available from: <http://www.beebase.org/> (accessed June 2008).

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., et al., 2009. Flybase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37, D555–D559.

Vilhelmsen, L., 2001. Phylogeny and classification of the extant basal lineages of the Hymenoptera (Insecta). Zool. J. Linn. Soc. 131, 393–442.

Vilhelmsen, L., 2003. Phylogeny and classification of the Orrusidae (Insecta: Hymenoptera), a basal parasitic wasp taxon. Zool. J. Linn. Soc. 139, 337–418.

Vilhelmsen, L., Mikó, I., Krogmann, L., 2010. Beyond the wasp-waist: structural diversity and phylogenetic significance of the mesosoma in apocritan wasps (Insecta: Hymenoptera). Zool. J. Linn. Soc..

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., et al., 2005. SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res. 33, D399–D402.

Wheeler, W.C., Whiting, M., Wheeler, Q.D., Carpenter, J.M., 2001. The phylogeny of the extant hexapod orders. Cladistics 17, 113–169.

Whitfield, J.B., 1992. Phylogeny of the non-aculeate Apocrita and the evolution of parasitism in the Hymenoptera. J. Hymenoptera Res. 1, 3–14.

Whitfield, J.B., 1998. Phylogeny and evolution of host–parasitoid interactions in Hymenoptera. Annu. Rev. Entomol. 43, 129–151.

Whitfield, J.B., Baker, M.R., Littlewood, 2003. Phylogenetic Insights into the Evolution of Parasitism in Hymenoptera. Adv. Parasitol. Academic Press, 69–100.

Whitfield, J.B., Cameron, S.A., Huson, D.H., Steel, M.A., 2008. Filtered z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. Syst. Biol. 57, 939–947.

Whiting, M.F., 2002. Phylogeny of the holometabolous insect orders: molecular evidence. Zool. Scr. 31, 3–15.

Wiegmann, B.M., Trautwein, M.D., Kim, J.-W., Cassel, B.K., Bertone, M.A., et al., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7, 34.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.