Blackwell Publishing

Cladistics 23 (2007) 130–144

Cladistics

10.1111/j.1096-0031.2006.00132.x

The nematode–arthropod clade revisited: phylogenomic analyses from ribosomal protein genes misled by shared evolutionary biases

Stuart J. Longhorn[1,2,3], Peter G. Foster[2] and Alfried P. Vogler[1,3]

[1]*Department of Entomology and* [2]*Department of Zoology, Natural History Museum, Cromwell Road, London, SW7 5BD, UK,* [3]*Division of Biology, Imperial College London, Silwood Park Campus, Ascot, SL5 7PY, UK*

Accepted 29 June 2006

Abstract

Phylogenetic analysis of major groups of Metazoa using genomic data tends to recover the sister relationships of arthropods and chordates, contradicting the proposed Ecdysozoa (the molting animals), which group the arthropods together with nematodes and relatives. Ribosomal protein genes have been a major data source in phylogenomic studies because they are readily detected as Expressed Sequence Tags (ESTs) due to their high transcription rates. Here we address the debate about the recovery of Ecdysozoa in genomic data by building a new matrix of carefully curated EST and genome sequences for 25 ribosomal protein genes of the small subunit, with focus on new insect sequences in addition to the Diptera sequences generally used to represent the arthropods. Individually, each ribosomal protein gene showed low phylogenetic signal, but in simultaneous analysis strong support emerged for many expected groups, with support increasing linearly with increased gene number. In agreement with most studies of metazoan relationships from genomic data, our analyses contradicted the Ecdysozoa (the putative sister relationship of arthropods and nematodes), and instead supported the affinity of arthropods with chordates. In addition, relationships among holometabolan insects resulted in an unlikely basal position for Diptera. To test for biases in the data that might produce an erroneous arthropod–chordate affinity we simulated sequence data on tree topologies with the alternative arthropod–nematode sister relationships, applying a model of amino acid sequence evolution estimated from the real data. Tree searches on these simulated data still revealed an arthropod–chordate grouping, i.e., the topologies used to simulate the data were not recovered correctly. This suggests that the arthropod–chordate relationships may be obtained erroneously also from the real data even if the alternative topology (Ecdysozoa) represents the true phylogeny. Whereas denser taxon sampling in the future may recover the Ecdysozoa, our analyses demonstrate that recent phylogenomic studies may be affected by as yet unspecified biases in amino acid sequence composition in the model organisms with available genomic data.

Molecular data have greatly changed the perspective on relationships of the Bilateria, i.e., the phyla of animals with bilateral symmetry developing from three germ layers (triploblasts). These data suggest that the traditional view of an acoel-like ancestor, which progressively acquired a coelome and differentiated internal organs, is invalid. The revised phylogenetic analyses moves the pseudocoelomate Nematoda away from the base of the bilaterian tree to a position near the Arthropoda with which they were grouped as the Ecdysozoa, the "molting animals" (Aguinaldo et al., 1997; see Fig. 1). The latter are separated from another major group of protostome phyla with spiral cleavage (Spiralia), which includes the Lophotrochozoa (Annelida, Mollusca, Brachipoda, etc.) and the formerly basal Rotifera and Plathyhelminthes (Adoutte et al., 2000). Initially, the revised taxonomic divisions were based on 18S rRNA evidence alone (Aguinaldo et al., 1997), but a growing suite of nuclear markers and combined morphological–molecular studies have continued to support these findings (Manuel et al., 2000; Peterson and Eernisse, 2001; Giribet, 2002; Ruiz-Trillo et al., 2002; Mallatt et al., 2004). Many of the changes, including the close relationships of arthropods and nematodes, had already been anticipated in the

*Corresponding author: *E-mail address*: apv@nhm.ac.uk

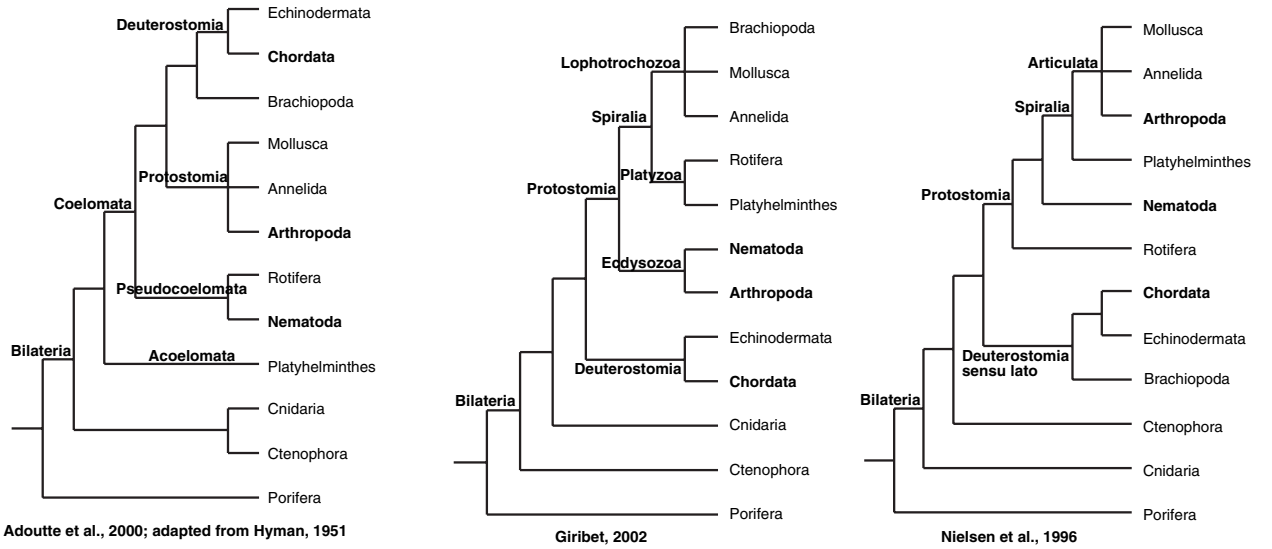

Fig. 1. Relationships of selected metazoan phyla. (A) Traditional viewpoint based on morphological data according to (Adoutte et al., 2000; adapted from Hyman, 1951). (B) The "new phylogeny" of Metazoa based on combined morphological and molecular data (Giribet, 2002). (C) Morphology based analysis prior to the formal establishment of Ecdysozoa (Nielsen et al., 1996). These trees were simplified by pruning terminals from the published topologies. Three taxa (Chordata, Arthropoda, Nematoda) critical for the debate about Coelomata are shown in bold.

morphological literature (Eernisse et al., 1992; Nielsen et al., 1996). However, controversy specifically about the validity of the Ecdysozoa remained based on morphological evidence (e.g., Nielsen, 2001; Scholtz, 2002), while the conclusions from 18S rRNA analyses were heavily criticized (Wägele et al., 1999) but defended by others (Zrzavy, 2001).

With the emergence of genome data for several animal species, deep metazoan relationships have become a test case for the use of such data in phylogenetics. Owing to the taxon sampling, initial studies have mainly focused on the validity of the Ecdysozoa. The surprising result was that the arthropods (*Drosophila melanogaster*) grouped with the representative of chordates (*Homo sapiens*) rather than nematodes (*Caenorhabditis elegans*), apparently refuting the Ecdysozoa (Mushegian et al., 1998; Hausdorf, 2000; Blair et al., 2002). This affinity of arthropods with chordates broadly reflects the traditional view of close relationships of taxa possessing a true body cavity, a coelom, which current literature refers to as the "Coelomata hypothesis". The arthropod–chordate (Coelomata) clade continues to be supported by studies using genome and Expressed Sequence Tags (ESTs), even when several hundred (Blair et al., 2005; Wolf et al., 2004; Philip et al., 2005) or several thousand (Copley et al., 2004; Dopazo et al., 2004) genes were used. The most recent study of all available eukaryotic genome sequences, selecting genes with well-established orthology, also supports this relationship (Ciccarelli et al., 2006). However, it has been argued that some of these studies are affected by inherent biases in genome evolution, such as consistent loss of genes in the nematode *C. elegans* (Copley et al., 2004) and idiosyncratic rates of sequence evolution (Telford, 2004; Dopazo and Dopazo, 2005), while phylogenetic inferences are further impeded by the presumed rapid diversification of major metazoan lineage (Rokas et al., 2005). The removal of the fastest evolving sites or genes also led to a reduction in support for the Coelomata (Philip et al., 2005; Philippe et al., 2005), possibly indicating the role of long-branch attraction (LBA) phenomena. Similarly, the exclusion of sites with different substitution rates between *D. melanogaster* and *C. elegans* (relative to humans) resulted in the first genomic scale data set that provided support for the Ecdysozoa (Dopazo and Dopazo, 2005). Recent studies that incorporate a broader taxonomic sampling of arthropods than Diptera alone, like inclusion of a representative Lepidoptera (silk worm, *Bombyx*) and Hymenoptera (honey bee, *Apis*) plus representatives of Spiralia, are in better agreement with the revised phylogenetic scheme, arguing against a basal position of nematodes (i.e., against the Coelomata). However, even with improved taxon sampling, phylogenetic analyses based on genomic data may not support a monophyletic Ecdysozoa due to presumed erroneous grouping of Nematoda with the Platyhelminthes (Philippe et al., 2004, 2005).

Another trend apparent with the advent of genome level analysis is the largely automated data compilation, where BLAST searches are used to retrieve similar sequences without evaluating orthology, based on automated alignments (Dopazo and Dopazo, 2005; Philip et al., 2005). While this may be acceptable practice with well-curated genome data, it may be less so using EST sequences due to the need for first

compiling full-length transcripts from multiple redundant ESTs, which frequently differ in sequence to various degrees. The difficulty of establishing gene orthology is a general problem of genome level data, and many loci fail in rigorous tests of orthology (Blair et al., 2005; Philip et al., 2005). Stringent data filtering to remove likely paralogs typically leads genome-based studies to use only a limited set of genes in phylogenetic analyses, e.g., only 31 orthologous genes from full genome sequences (Ciccarelli et al., 2006).

Here we address questions about confounding factors that might prevent the recovery of an arthropod–nematode sister relationship (Ecdysozoa). A possible problem with recent data sets is the use of Diptera as representative arthropods (using only *D. melanogaster* plus occasionally *Anopheles gambiae*), given that dipteran genomes display an unusual shift in substitution rate relative to other insects (Friedrich and Tautz, 1997; Hwang et al., 1998; Savard et al., 2006). With newly available EST data for a broader spectrum of insect orders (Hughes et al., 2006), this limitation can now be partly overcome, permitting an investigation of how metazoan relationships might be misled by using Diptera as the sole representative of arthropods. We were also interested in the differences in phylogenetic signal that each gene would contribute, and how support would depend on the number of genes used.

For these analyses, we carefully curated sequences for 25 ribosomal protein (RP) genes of the small ribosomal subunit for three major metazoan lineages, Arthropoda, Nematoda and Deuterostomia (Chordata and Urochordata). The RP genes are conserved at the amino acid level relative to most other nuclear genes (Brochier et al., 2002), which facilitates sequence alignment and homology assignment between disparate taxa. RPs are also among the few genes without evident paralogy in most metazoan genomes (Uechi et al., 2001; Karsi et al., 2002; Landais et al., 2003; Hughes et al., 2006). Hence, these data are appropriate for investigating pertinent problems of the use of genomic sequences in phylogenetics, including issues of data quality, taxon sampling, and possible biases in sequence composition that confound the phylogenetic signal. Using data simulations, we show that the character variation in these sequences makes it unlikely that the arthropod–nematode sister relationship (Ecdysozoa) is recovered with the available data, even if this group were truly monophyletic.

Materials and methods

Generating well curated RP data sets

Figure 2 shows the taxa used in this analysis and their classification. Public databases already contain DNA sequences assembled from ESTs to delimit open reading frames (ORFs) and generate putative protein sequences. RP sequences for the following taxa were first obtained from the non-redundant (nr) protein database of GenBank: *H. sapiens* (Uechi et al., 2001), *D. melanogaster*, *C. elegans*, *Ictalurus punctatus* (Vertebrata, Pisces) (Karsi et al., 2002), *Spodoptera frugiperda* (Lepidoptera) (Landais et al., 2003), *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Fungi). To increase taxon sampling and verify whether nr database sequences matched those from ESTs, we exploited clusters of ESTs in the TIGR Gene Indices database (Quackenbush et al., 2001) for 10 vertebrates, one urochordate, three nematodes and three insects (Diptera and Hymenoptera). For each gene, we evaluated each cluster of ESTs produced by automated procedures of Gene Indices. For each cluster identified by similarity to a given RP gene, we established a consensus sequence, using the consistency of base calls across clones as an indicator of base confidence. EST clustering may generate multiple groups of highly similar sequences even when only a single gene copy exists, whereby artifactual secondary clusters may be based on ESTs of low read quality, or include spurious ESTs with unspliced introns (Liang et al., 2000). When alternative EST clusters were found in the TIGR assemblies, the number of constituent clones, presence of long non-coding regions, and inclusion of specific motifs were used to choose the best assemblies for each gene, discarding spurious assemblies with low numbers of sequences (< 10) unless none other was available. Usually a single cluster dominated for each gene, for which the coding sequences were identified using Framefinder (Slater, 2000). For taxa with sequences in both the nr database (including RefSeq; Pruitt et al., 2000), and the TIGR Gene Indices databases, we compared sequences using BLAST to identify agreement between curated nr data and consensus ESTs. Accessions of sequences in the nr database that match our consensus ESTs are given in the supplementary information. Additional RP sequences were obtained from recently curated EST libraries of Coleoptera (beetles) and Lepidoptera (butterflies) (Hughes et al., 2006). For Coleoptera, individual libraries had missing data and were combined for Adephaga (*Melademа coriacea*, *Carabus granulatus*, *Cicindela campestris* or *C. littoralis*); Polyphaga, Elateriformia (*Dascillus cervinus* or *Julodis onopordi*); and Cucujiformia (*Platystomos albinus*, *Curculio glandium*, *Timarcha balearica* or *Biphyllus lunatus*).

Protein sequences were aligned using ClustalX (Thompson et al., 1997), including data from GenBank and new conceptual EST translations when both were available. This ensured EST data could be used to generate high-quality amino acid sequences, and established which genes might be confounded by paralogs. Neighbor-joining (NJ) analyses in PAUP* 4.0b10 (Swofford, 2002) generally identified a single

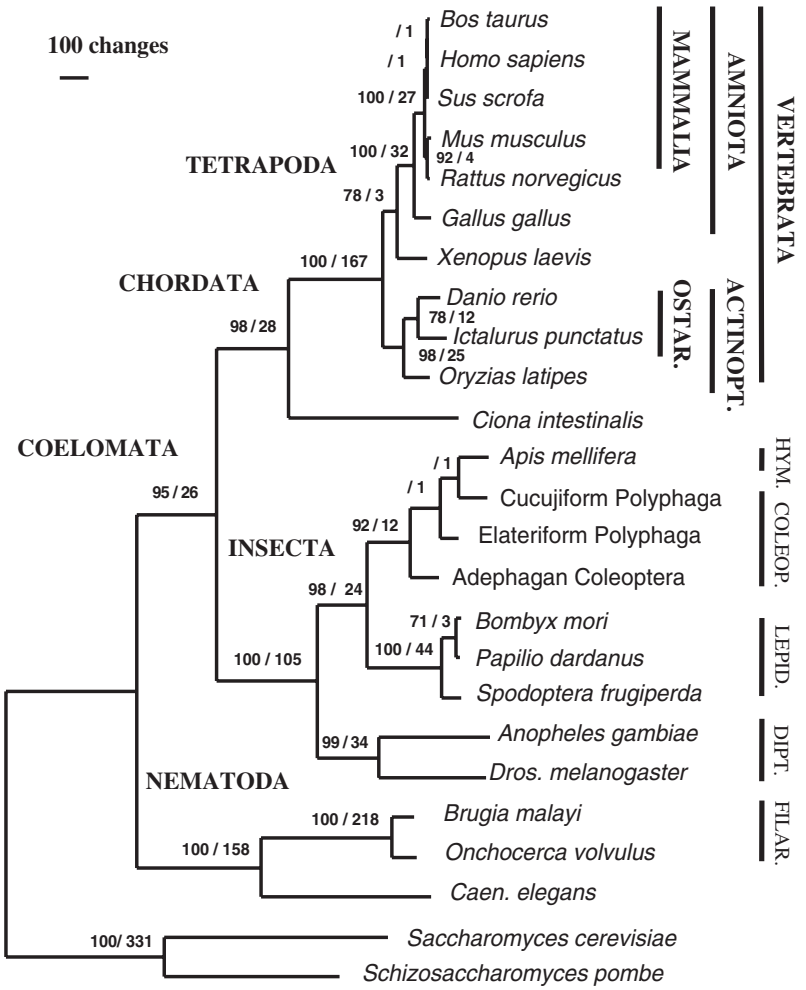

Fig. 2. Topology obtained by simultaneous analysis of 25 RP genes using parsimony. Bootstrap and PBS are given for each node. ACTINOPT, Actinopterygii; OSTAR, Ostariophysi; HYM, Hymenoptera; COLEOP, Coleoptera; LEPID, Lepidoptera; DIPT, Diptera; FILAR, Filarioidea.

transcript for each taxon with greatest EST support, least ambiguity, and longest ORF. If, for a particular taxon, we found multiple amino acid sequences each supported by equal numbers of ESTs, the copy with lower divergence from other taxa was retained, provided these putative paralogs clustered together in NJ analyses. If alternative forms did not cluster together, affected sequences were excluded, and if multiple taxa were confounded by alternative amino acid sequences the entire gene was excluded from further analysis. We also excluded the proteins S0/A, S16e, S21e and S29e from this study because of limited data availability.

Phylogenetic analyses

Phylogenetic analyses to assess congruence between remaining genes were initially conducted using heuristic parsimony searches on a concatenated matrix of amino acids, with insertion/deletions (indels) coded as 21st state to maximize informative character transformations. To assess nodal support from different genes (partitions), we used partitioned branch support (PBS) (Baker and DeSalle, 1997) using constraints files generated with TreeRot 2c (Sorenson, 1999) and PAUP. PBS measures the support from each partition on the simultaneous analysis topology, showing relative contributions to the branch support at each node. The PBS values across all nodes in a cladogram (sum PBS) equal the total branch support (sum BS) of the combined analysis. Partitioned hidden branch support (PHBS) is the difference between branch support on the simultaneous analysis topology (PBS) and unconstrained analyses (BS) for each node and gene (Gatesy et al., 1999). The PHBS is a positive value when branch support for a partition is increased under simultaneous analysis over that in separate analysis, and negative if relative support levels decrease. For each gene, PBS was normalized for the number of character steps under simultaneous analysis to provide a relative measure of support. The effect of including alignment variable

regions was investigated by repeating parsimony searches on a second alignment with variable regions (columns containing indels) excluded. This second alignment is deposited at EMBL as ALIGN_000747. Alternative topologies were compared using Templeton's (1983) tests in PAUP.

For model-based phylogenetic analyses, we assumed a homogeneous model across genes and taxa. Using the second alignment excluding variable regions, we evaluated different model parameters (substitution matrix, proportion of invariant sites, shape of the gamma distribution) in TREE-PUZZLE 5.2 (Schmidt et al., 2002) using empirical base frequencies. From available models, the WAG matrix (Whelan and Goldman, 2001) was favored by the Akaike Information Criterion, describing among-site rate variation with a gamma distribution across four categories (WAG + Γ). Accounting for invariant sites did not lead to a significant improvement in model fit. Pair-wise distances from the preferred model were generated by TREE-PUZZLE and used for building NJ trees in NEIGHBOR (Felsenstein, 1989), with bootstrap support (500 replicates) from PUZZLE-BOOT (A. Roger and M. Holder, unpublished). We used the same model for Bayesian analyses in MrBayes 3.0b4 (Huelsenbeck and Ronquist, 2001). For the latter, the first 200 samples were discarded, and convergence assessed in three ways. First, we plotted the likelihood of the chain to show a plateau. Second, two independent runs were performed from different random starting trees. Third, post burn-in samples were divided into three parts and a consensus tree was made from each to ensure convergence of the topology. ML trees were obtained with PHYML 2.4.4 (Guindon and Gascuel, 2003) using the WAG + Γ model, and bootstrap analysis was performed with 1000 replicates using the same model and estimating among-site rate variation for each pseudoreplicate. Composition tests and simulations of sequence data on predefined tree topologies were made using p4 (Foster, 2004). Likelihoods of contending topologies were assessed with CONSEL (Shimodaira and Hasegawa, 2001) using Shimodaira–Hasegawa (SH) and Approximately Unbiased (AU) tests (Shimodaira, 2002).

Results

The data matrix included 24 small subunit RP genes that had a putative ortholog present in at least 20 of the 25 taxa. We also included acidic RPP0 for comparison, which shows an unusual sequence composition rich in alanine (A) and lysine (K). As had already been noted in the literature (Karsi et al., 2002; Zhang et al., 2002), probable paralogs were found for S27e and alternative splice products were obtained in S24e of vertebrates (each supported by multiple ESTs), whereas mammals had alternative forms for S4e. Therefore, we excluded genes S4e, S6e, S24e and S27e due to expressed paralogs or multiple splice forms. Recent duplications of RPs were common in *S. pombe*, *S. cerevisiae*, and the amphibian *X. laevis*, but in each gene affected, paralogs were more similar to one another than to other taxa, and hence the copy with least divergence was used in phylogenetic analyses. Paralogous sequences in some genes (e.g., S14eA and B in Diptera), were also either identical or near identical at the protein level and, where divergent, choice between alternate forms had no effect on tree topologies.

For taxa with well-curated RP sequences available in the nr database (mainly *H. sapiens*, *D. melanogaster* and *C. elegans*), the annotated coding regions matched our corresponding consensus EST sequences in every gene whether comparing protein or DNA sequences. For other taxa, where curated RP sequences in GenBank were mostly from automated procedures (such as *M. musculus* and *A. gambiae*), exact matches with consensus ESTs were generally only at the protein level, in which case we preferred our EST-based annotations. Consensus ESTs often predicted the limits of the coding regions better than automated RefSeq annotations, with greater conservation of reading frames from EST data.

Individual gene analyses

The size of amino acid data matrices and basic tree statistics from parsimony analyses are given in Table 1 for 25 RP genes. Cases of length variation (S2e, S10e) were often restricted to the terminal regions of proteins. Individually, parsimony analyses of each RP showed low resolution and generally weak nodal support (Supplementary material Fig. S1). In Chordata, several well-established taxonomic groups (e.g., Meyer and Zardoya, 2003) were recovered frequently, including the Chordata (recovered from 14 proteins), Vertebrata (24), Mammalia (10), Amniota (11), Actinopterygii (13) and Ostariophysi (seven). In contrast, few protein sequences supported the Rodentia (three) or Tetrapoda (four). The monophyly of insects was supported by the majority of proteins (21), as was the monophyly of the Lepidoptera (23, with remaining genes present for just a single taxon). Less than half of the genes (11) supported the monophyly of Diptera. None of the individual gene partitions supported the monophyly of the Coleoptera, although gene sampling was less complete for this order with only four genes present across all three lineages. At basal nodes, all genes (25) supported the monophyly of the Metazoa relative to fungi, and the monophyly of the Nematoda. Nine proteins recovered a node joining chordates and insects in accordance with the Coelomata, while only five grouped nematodes with insects in support of the Ecdysozoa. In addition, two proteins grouped nematodes and chordates to the exclusion of both insects and fungi.

Table 1
Characteristics of single gene phylogenies and influence of simultaneous analyses (SA)

Partition (gene)	Amino acid sites	Gapped sites (%)	Parsimony informative sites	Individual protein analyses (Ind.) versus SA								Hidden support (PHBS)	PBS/tree length
				Consistency Index		Retention Index		Min. steps		Branch support			
				Ind.	S.A	Ind.	SA	Ind.	SA	Ind.	SA		
P0	320	6.6	182	0.65	0.63	0.73	0.71	830	843	98	93.6	−4.4	0.11
S2e	299	20.0	144	0.79	0.76	0.81	0.80	615	620	87	79.2	−7.8	0.13
S3e	263	11.4	109	0.82	0.78	0.76	0.73	501	510	65	46.3	−18.7	0.09
S3Ae	272	10.6	140	0.75	0.75	0.78	0.78	634	635	85	77.3	−7.7	0.12
S5e	233	16.3	83	0.81	0.80	0.77	0.75	413	419	50	51.3	1.3	0.12
S7e	201	7.5	111	0.75	0.75	0.80	0.80	439	439	56	78	22	0.18
S8e	210	5.7	104	0.67	0.65	0.71	0.69	440	450	59	66.3	7.3	0.15
S9e	197	8.1	63	0.69	0.68	0.74	0.73	264	265	32	25.5	−6.5	0.10
S10e	172	42.4	105	0.76	0.75	0.81	0.80	538	544	90	94	4	0.17
S11e	165	12.1	71	0.73	0.73	0.74	0.74	318	319	29	39	10	0.12
S12e	157	18.5	98	0.78	0.76	0.77	0.75	398	399	55	32	−23	0.08
S13e	151	0	50	0.69	0.68	0.74	0.72	181	184	29	29	0	0.16
S14e	152	9.9	46	0.76	0.73	0.83	0.81	156	160	24	24.8	0.8	0.16
S15e	153	9.2	59	0.80	0.79	0.83	0.81	258	262	25	21	−0.4	0.08
S15Ae	130	0.8	36	0.68	0.66	0.74	0.72	127	130	16	19	3	0.15
S17e	145	16.5	80	0.80	0.77	0.85	0.82	298	307	65	56.8	−8.3	0.19
S18e	156	5.1	58	0.66	0.63	0.81	0.79	211	217	36	40.7	4.7	0.19
S19e	162	12.3	106	0.73	0.73	0.76	0.76	467	467	62	76.5	14.5	0.16
S20e	124	8.9	54	0.84	0.78	0.85	0.82	238	246	49	52.5	3.5	0.21
S23e	145	1.4	39	0.79	0.77	0.88	0.87	112	114	26	30.7	4.7	0.27
S25e	130	32.3	74	0.81	0.79	0.87	0.85	284	293	60	50	−10	0.16
S26e	124	9.7	56	0.81	0.78	0.81	0.78	254	260	34	42	8	0.16
S27Ae	169	15.4	40	0.88	0.87	0.92	0.91	192	194	39	36.5	−2.5	0.19
S28e	70	8.6	29	0.81	0.77	0.90	0.87	75	79	22	17	−0.5	0.22
S30e	141	58.2	101	0.79	0.78	0.78	0.76	503	512	75	78.2	8.7	0.15
Sum	4441	–	2038	–		–		8744	8868	1268	1257	−11	
Aver.	177.6	13.9	81.5	0.76	0.74	0.80	0.78	349.8	354.7	50.2	50.7	–	0.15

The effect of simultaneous analysis

Simultaneous analysis of 4441 amino acids and 2038 parsimony informative sites produced a single tree of 8868 steps (CI = 0.74, RI = 0.76; Fig. 2). As with individual analyses, relationships within chordates conformed to the well-established topology, with the only exception that relationships within mammals were weakly supported (low BS) presumably due to low sequence variation across mammals. At basal metazoan nodes, simultaneous analysis favored an arrangement consistent with the Coelomata with relatively strong overall branch support (BS = 26). However, there was little co-ordinate behavior among partitions at this node (Fig. 3), as different genes produced conflicting signal, and many partitions supported alternative relationships (e.g., S20e, S25e, etc.). Unlike most other nodes, the sum of branch support from the individual gene analyses was negative here (BS = −6). Hence, the recovery of the Coelomata topology was largely due to hidden support occurring at a high level of PHBS = 32. Between-partition conflict was lower at other nodes, and in general, support levels increased under simultaneous analysis (positive PHBS across genes), with the exception of the Tetrapoda (Mammalia + *G. gallus* + *X. laevis*), which was the only node within chordates to show hidden conflict (negative PHBS). Other deeper nodes, such as the Filarioidea, Nematoda and Metazoa showed positive support from most partitions under both individual and simultaneous analysis, with a substantial level of hidden branch support (positive PHBS). At the Insecta node, support levels were more erratic, perhaps influenced by missing data in Coleoptera, *P. dardanus*, and *A. mellifera*. The only nodes within insects to show overall hidden conflict from simultaneous analysis were the Lepidoptera with high negative PHBS = −75, and PHBS = −6 for the group of *B. mori* plus *P. dardanus*. This level of conflict was surprising, because partitions with complete data for all taxa (S3e, S20e, S28e) showed strong positive support for the Lepidoptera under individual analyses, although they disagreed about the relationships within the order.

The amino acid sequences of individual RPs ranged from 70 to 320 residues in length, which correlated well with the number of parsimony informative sites ($Y = 0.47x$; $R^2 = 0.67$) and number of steps ($Y = 0.20x$; $R^2 = 0.41$) (Table 1), although the latter differed by a factor of more than 10 between genes. There was a slight

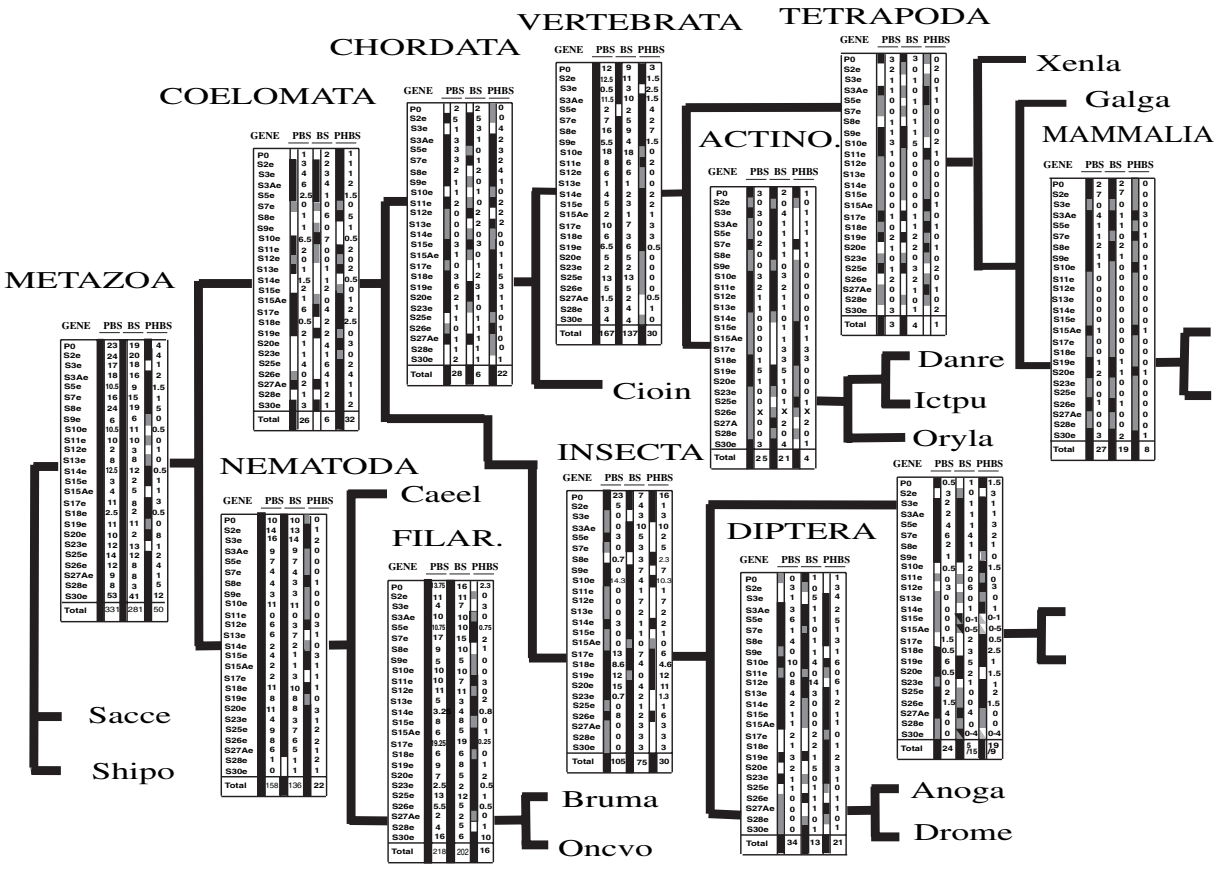

Fig. 3. Branch support from simultaneous versus individual gene analyses. At each node PBS, BS, and PHBS scores are given, shown for selected nodes of the tree from Fig. 2. Following the color coding of Gatesy et al. (1999), positive values are indicated by black boxes to the left, negative values by white boxes and zero values by gray boxes. Ambiguous scores due to missing data are given as their range (e.g., 0–1), missing values as X. Note the preponderance of hidden support (PHBS) defining the Coelomata node.

negative correlation of both the ensemble CI and RI with alignment length ($Y = -0.0003x$, $R^2 = 0.06$; and $Y = -0.0004x$; $R^2 = 0.21$, respectively), indicating that character variation in the smaller proteins was slightly less homoplastic (Table 1). Importantly, sequence length showed a positive correlation with overall levels of nodal support. Evaluation of sequence length (number of amino acids) against branch support in separate analysis (sum BS) or simultaneous analysis (sum PBS) yielded similar results (Fig. 4), and revealed greater nodal support in larger partitions ($Y = 0.28x$, $R^2 = 0.43$; and $Y = 0.28x$, $R^2 = 0.31$, respectively). However, hidden support generally arose from smaller partitions ($Y = -0.03x$; $R^2 = 0.04$, see Table 1), while many larger ones actually gave negative PHBS values, though the regression fit between sequence length and PHBS was weak. Taken together, these results suggest that larger partitions provide much of the branch support for simultaneous analysis and influence smaller partitions, even though these smaller partitions showed character behavior that is more consistent among sites (higher CI, RI, etc.). Finally, as it has been suggested that measures of branch support using BS (and derivative measures) are not necessarily comparable among partitions, we evaluated the variation in PBS across partitions by normalizing by tree length. Interestingly, this now revealed a negative relationship between gene length and "corrected" PBS ($Y = -0.0004x$; $R^2 = 0.26$, see Table 1), indicating that the smallest partitions (like S23e, and S28e) provided greater support in the simultaneous analysis relative to their size, even though larger partitions do provide most nodal support overall.

When the interaction between partition size and branch support was investigated at particular nodes, focusing on four key groups (Metazoa, Coelomata, Insecta, Vertebrata), the overall trend of increased nodal support from simultaneous analysis also held up broadly, although data scatter was large and affected by negative support values for individual genes (Fig. 5). Correlation of branch support and number of amino acid residues showed that the increase from simultaneous analysis is very similar for all four nodes

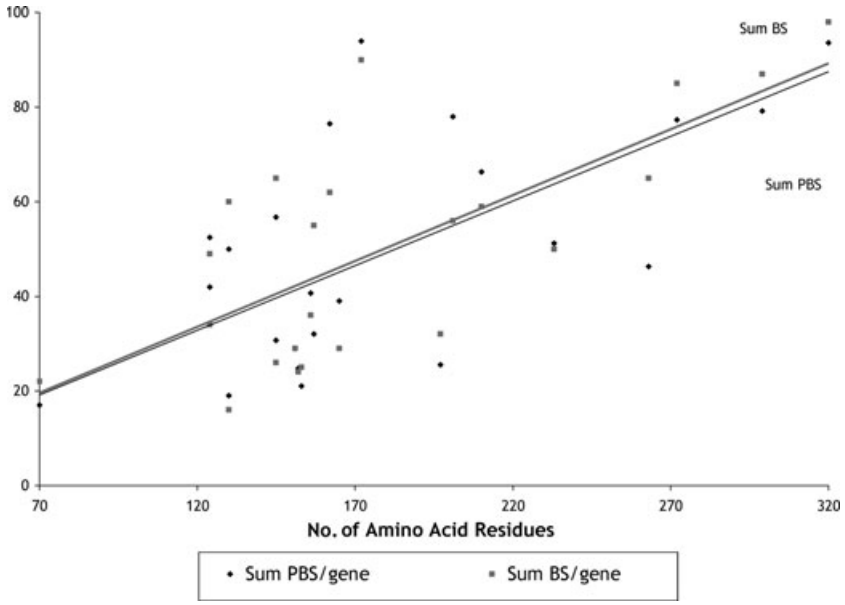

Fig. 4. Total branch support across RP topologies under individual (Sum BS) or simultaneous (Sum PBS) analysis plotted against the number of amino acid residues. All sites including gapped positions of the alignments were included in this analysis.

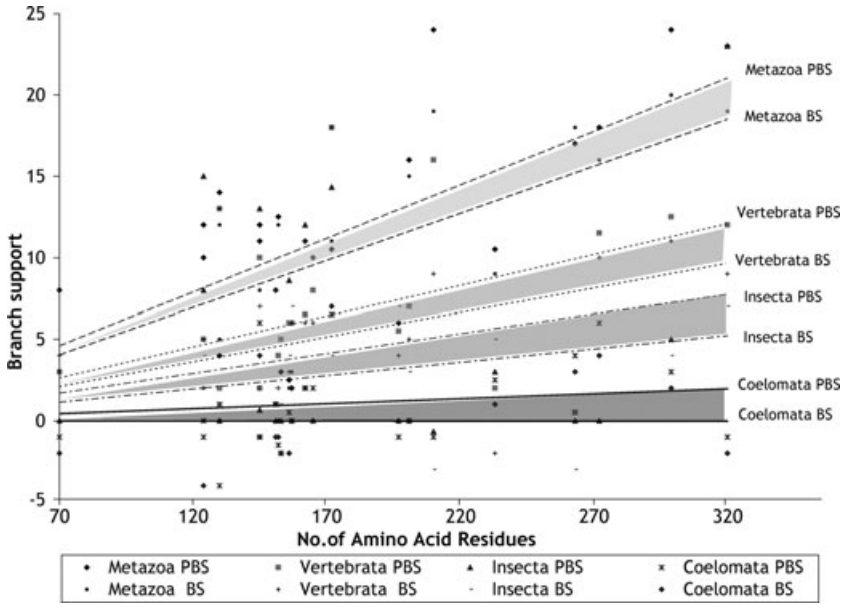

Fig. 5. Correlation of branch support and the number of amino acid residues in each gene at four nodes of the phylogeny shown for individual (BS) and simultaneous (PBS) analyses. The straight lines show the regression for the BS and PBS values (respectively) at the four test nodes. Shaded areas represent the gain in support levels from hidden support in the simultaneous analysis.

investigated, including the Coelomata node where the regression slope of the separate analysis was essentially zero, indicating support from simultaneous analysis only. At these nodes, the greatest discrepancy in support level was between individual and simultaneous analysis of larger partitions. The discrepancy in separate and simultaneous analysis also seemed to be dependent on the number of amino acids in a partition. For example, unlike the results from overall support, the largest partitions showed hidden support in simultaneous analysis at these nodes, while the smallest partitions showed increased support to a lesser degree. These

results would suggest that the topology of small proteins did not fit well the overall phylogenetic signal, perhaps due to greater functional constraints on a shorter amino acid sequence.

Robustness of the Coelomata topology

The simultaneous analysis topology (Fig. 2) was unexpected with regard to the strongly supported basal (i.e., sister to all other Holometabola) position of the Diptera among insects. Equally, there was strong support for the affinity of the chordates to insects rather than nematodes as would have been expected under the Ecdysozoa hypothesis. When indel containing columns were removed from the matrix to reduce potential problems of alignment ambiguity, the data included 3833 amino acid residues, of which 1561 were parsimony informative. Simultaneous parsimony analysis returned two shortest trees of 6400 steps (CI = 0.72, RI = 0.77) whose topology and branch support were very similar to the tree obtained from the indel containing data matrix (Fig. 6A). This suggests that misalignment of sites should not be implicated as a major confounding factor affecting topologies. The insect–chordate clade remained strongly supported (BS = 25), while the alternative insect–nematode relationship (Ecdysozoa) required an extra 25 steps, and was rejected as a significantly worse topology with a Templeton test ($P < 0.01$).

Equally, model-based phylogenetic procedures were used to complement the parsimony analysis, as they may react differently to long-branch effects. All methods used, including NJ (Fig. 6B), Bayesian and ML (Fig. 6C), consistently recovered topologies almost identical to the unconstrained parsimony trees, again favoring the insect–chordate clade. However, probabilistic tests of alternative topologies were less dismissive of the Ecdysozoa. For example, the preferred NJ tree from model-based distances (Fig. 6B), was not statistically favored over an alternative topology that enforced the Ecdysozoa (SH, $P = 0.56$; AU, $P = 0.27$).

As the Ecdysozoa remains a plausible alternative for metazoan relationships, we tested for factors that could have biased against recovery of this group. First, we examined the possible effect of amino acid heterogeneity on the topology, as similar sequence composition among taxa may cause erroneous groupings even at the amino acid level (Foster and Hickey, 1999). Using an unadjusted null frequency distribution, the standard chi-square test (using the p4 software) accepted the data set composition as homogeneous ($P = 1.000$), even after removal of constant sites ($P = 0.99$). However, it is well known that this homogeneity test does not take into account relatedness of the sequences, and suffers from a high probability of type II error (i.e., the false null hypothesis of compositional homogeneity is not rejected). In an alternative test, composition was re-evaluated

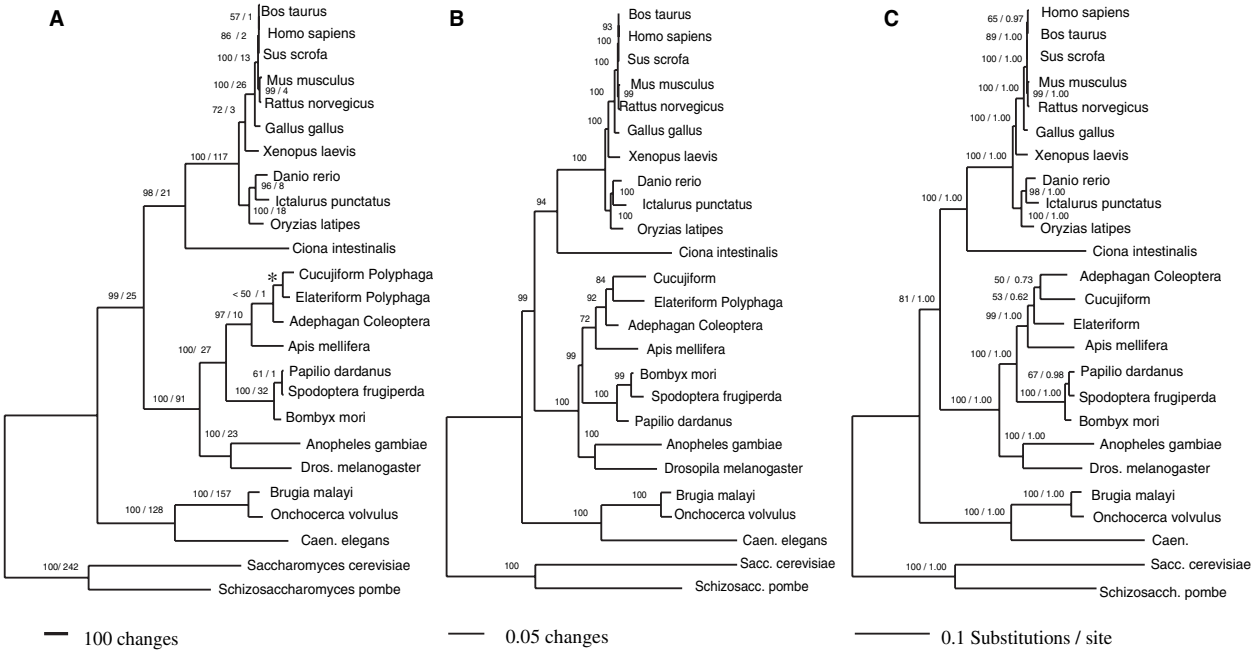

Fig. 6. Simultaneous analysis of 25 RPs based on indel-free positions. (A) One of two equally parsimonious trees, with bootstrap scores and BS given for each branch; the asterisk indicates the single node unresolved in the strict consensus. (B) Neighbor-Joining tree using model-based distances with bootstrap support. (C) Tree obtained from Bayesian inference; the topology shown was nearly to the ML tree except for the placement of *Apis* (inside of the basal Coleoptera node). Numbers on branches show bootstrap support from ML followed by posterior probabilities from Bayesian analysis.

using a null distribution of chi-square statistics from data simulated on the NJ tree using empirical base frequencies, optimized model parameters and branch lengths (see Foster, 2004). In this analysis, homogeneity was rejected ($P < 0.005$), indicating that amino acid frequencies do vary significantly across the tree, and that fungal sequences show the most deviant composition from the mean.

We then asked whether the compositional similarity (as the result of shared ancestry or convergence) could explain the recovery of the contentious nodes. To evaluate the effects of base composition on tree topology, we generated 200 non-parametric bootstrap pseudo-replicates from the original data, and for each calculated euclidean distances between sequence compositions (Lockhart et al., 1994). The consensus distance tree from NJ analysis (Fig. 7) did not show attraction between the nematodes and the outgroups (which would explain a placement of nematodes outside the insects and chordate clade). Neither did this tree show a placement of Diptera in the vicinity of any non-insect group which could be responsible for the basal position of *D. melanogaster* and *A. gambiae* in previous simultaneous analyses. This suggests that similarity of amino acid composition cannot be easily invoked to explain the major features of the tree topology.

There may be other biases in the sequence variation favoring the insect–chordate clade, including LBA, which is well known in parsimony analysis. (Long branch effects can also be problematic in model-based methods when the model does not fit well.) Long branches in the tree may go together because of common ancestry or due to artifacts from LBA. To see whether LBA is a plausible explanation, one can simulate data based on an alternative tree where the long branches are separate from each other in the simulation tree, and then perform a phylogenetic analysis on the resulting simulated data. If the recovered tree is similar to the simulation tree then that argues that it is robust to LBA, but if the long branches end up together then LBA may well be at play, and may have driven the branches together using the original data as well. Doing many replicates can give a measure of support for any trends observed. In this framework we simulated data based on the alternative Ecdysozoa tree, again modeled with parameters estimated from the original data (WAG matrix, gamma4, $\alpha = 0.53$, pInvar = 0.0) and using ML branch lengths. Parsimony analysis on this simulated data set resulted in a topology similar to the original tree from real data with support for the basal position of Nematoda and the insect–chordate relationship in 56% of the replicates, despite the fact that data were simulated on the Ecdysozoa tree. This suggests that the basal placement of the nematodes may indeed have been due to LBA. Although the Coelomata tree that we recovered may be correct, based on these simulations it is plausible that our tree could reflect the Ecdysozoa topology distorted by LBA.

In a similar way, we looked at the possibility that LBA could force the Diptera towards a basal position in the insects. Simulated data were generated on trees constrained for Diptera in a derived position as sister to Lepidoptera, while deeper nodes were additionally constrained to conform to either insects + chordates or insects + nematodes. Parsimony searches of data simulated on either of these trees now placed Lepidoptera basal in the insects with high support (83% and 80% of replicates, respectively), although Lepidoptera were not basal in the trees on which data were simulated. These results suggest that while long branches may be influencing the insect relationships, there is no clear inherent bias in character variation that would have resulted in the basal position of Diptera due to spurious attraction of the Diptera and non-insect taxa.

Discussion

The largely uniform support for the sister relationship of arthropods and chordates (or Deuterostomia) in genomic studies has been puzzling, and contradicts the "new bilaterian phylogeny" of recent years that suggested two major protostome lineages, Ecdysozoa and Lophotrochozoa/Spiralia, to the exclusion of Deuterostomia (including Chordates). The simultaneous analysis of 25 RP genes also supported the affinity of insects with chordates to the exclusion of nematodes even though insects were sampled in more detail than previous studies. This was not only evident in parsimony approaches, but equally using model-based approaches (Fig. 6A–C). It has been suspected that the elevated substitution rate in some nematodes confounds phylogenetic analyses of metazoan relationships, but previous data simulations did not confirm that LBA directly affects the position of nematodes (Wolf et al., 2004). In contrast, our simulations showed that data generated on an Ecdysozoa tree (under a model of rate variation whose estimation was independent of a particular tree topology) still produced a topology grouping insects and chordates, although the insects–nematodes arrangement is the true topology in these simulations. Equally, the ML tree obtained under JTT model used by Wolf et al. (2004) was topologically identical with our best ML tree, with slightly worse likelihood scores (–LogL = 45330.61 versus 45267.38 using the WAG model). This indicates that the insect–chordate relationship may indeed be an artifact of LBA among major lineages of Metazoa.

The precise cause for this erroneous result of the data simulation remains open, demonstrating the difficulty to establish if the Ecdysozoa are indeed a true clade or not. With these data, it is not likely that amino acid composition biases are responsible because the NJ

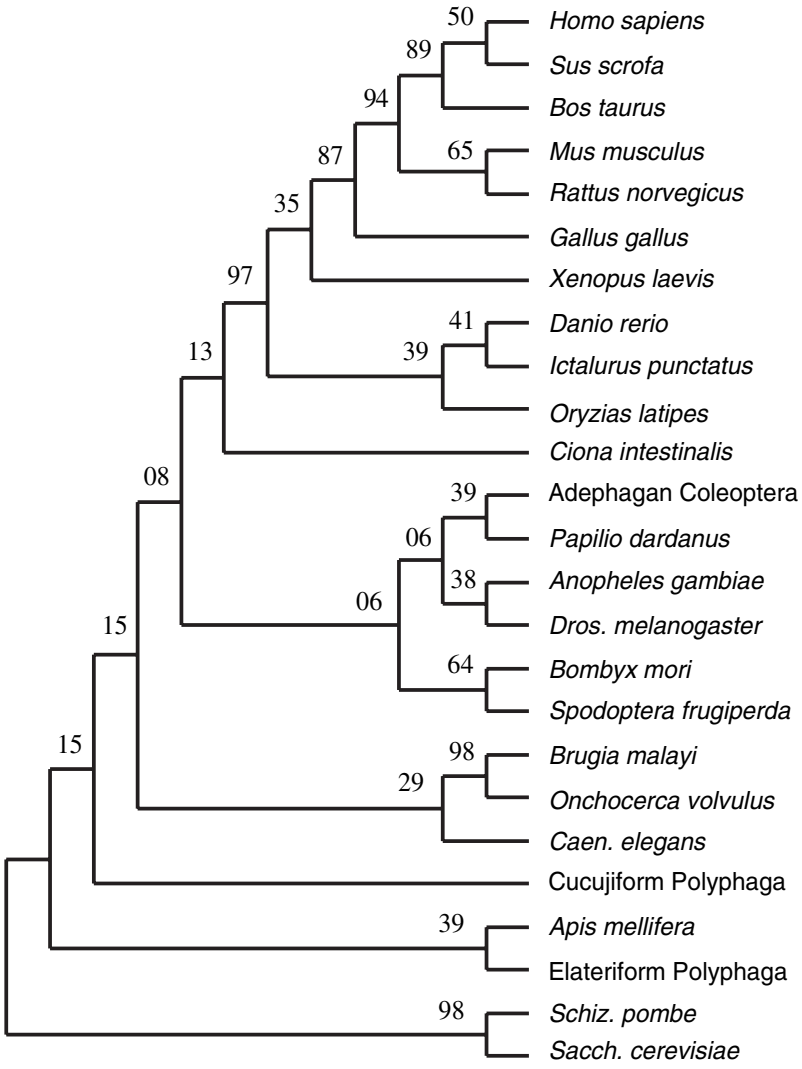

Fig. 7. Neighbor-joining tree based on pair-wise Euclidean distances inferred from the amino acid compositions of 27 RPs of a wide sample of Metazoan taxa. Numbers above nodes are the proportion of trees recovering these nodes in 200 bootstrap replicates of the amino acid data.

analysis on distances derived from compositional data did not provide any clear support for the insect–chordate topology (Fig. 7). We also established that the recovery of the insect–chordate clade is not clearly correlated with unusual properties of dipteran molecular evolution, including their faster rates of molecular substitution compared with other insects (Friedrich and Tautz, 1997; Savard et al., 2006). This was of particular concern because of the peculiar position of the Diptera as the sister group of all other holometabolan insects, which may be an indication of long-branch artifacts of Diptera with taxa outside insects, such as the chordates. However, data simulations on a tree that constrained Diptera to a more probable placement near Lepidoptera produced a signal that no longer puts Diptera as sister to all other insects, indicating that the sequence variation per se was not responsible for this spurious placement in this instance. Nevertheless, the insects remained allied with chordates in these simulations, showing that the biases that may inhibit the recovery of the insect–nematode relationships are not a result of specific features of the Diptera sequences. While the basal position of Diptera in unconstrained analyses may well be indicative of their faster rates, any bias that leads to the erroneous recovery of Coelomata in the simulated data is also shared by sequences from other insects.

Among the more striking observations of our analysis was the almost linear increase of Bremer support values with the greater number of genes and the number of sites per gene. While not all individual gene trees support the insect–chordate topology, under combined analysis RPs act synergistically to reveal ''hidden branch support'' (Gatesy et al., 1999) in favor of this clade. This result is

interesting in light of the genome-based study of Wolf et al. (2004) across six species of eukaryotes that analyzed proteins separately by functional categories. Unlike most cellular components, which strongly supported the Coelomata topology, their simultaneous analysis of RPs supported the Ecdysozoa when analyzed under likelihood. A re-analysis of our data with the model of Wolf et al. (2004) did not change the tree topology from that favoring the Coelomata, either using NJ or ML (Fig. 6B,C). This may be indication of how phylogenetic analyses of RPs might support either Ecdysozoa or Coelomata depending on different taxon sampling. The currently most complete taxon sampling for Metazoa (Philippe et al., 2004, 2005), dominated by (but not restricted to) RP genes, appears to produce trees that recover both the Ecdysozoa and the Lophotrochozoa under certain parameter choices. This might suggest that denser taxon sampling will ultimately solve the question in favor of the Ecdysozoa, although the tendency for a very unexpected grouping of Nematoda with Platyhelminthes (Philippe et al., 2005) suggests new complications from LBA with inferring relationships when certain taxa are added.

Beyond issues of taxon choice, approaches to compiling data for large-scale phylogenomic analyses remain rather crude. To rapidly increase the taxon sampling with current DNA technologies, highly expressed genes such as RPs from EST data will have to be the main source of data. Based entirely on ESTs, the curation in our current study could be used to rapidly identify high fidelity transcripts within a set of error-prone primary sequences. For many taxa, the problem of error-prone sequences is increasingly affecting sequences in the nr databases, where repeated submission of different conceptual translations, or un-curated automated assemblies have led to divergent entries for protein sequences. Divergent, near-redundant sequences can either be an artifact of low fidelity sequencing, frameshift errors, or represent true paralogs or differentially spliced transcripts. Choosing between these multiple accessions in public sequence databases can be difficult. The manually curated RefSeq accessions (Pruitt et al., 2000) are an attempt to identify sequences of highest read quality but are currently limited to a narrow range of taxa. By adopting the consensus approach, clustering ESTs and comparing the consensus to other precurated sequences, we could rapidly generate RP transcripts for additional taxa. Furthermore, consensus EST sequences were frequently identical to existing curated GenBank (RefSeq) accessions when available (e.g., in *H. sapiens* and *D. melanogaster*). We conclude that consensus ESTs can provide data of high quality, and even be used to discriminate between multiple primary sequences present in the curated databases. Finally, careful curation of each gene within our wider multigene data set greatly increased the confidence in both the sequences used here and the resulting trees.

Another problem with the use of genomic data is the potential abundance of pseudogenes and functional paralogs. In vertebrates, RP pseudogenes are common (Zhang et al., 2002), although these mainly cause problems for compiling paralogy free alignments when using genome sequences rather than EST data, which are unlikely to include non-transcribed pseudogenes. We found little evidence of functional duplications of metazoan RP genes, with the exceptions of S4e and S27e where paralogs were already known (Uechi et al., 2001; Karsi et al., 2002; Landais et al., 2003). Information about taxa with duplicated RP genes was refined in this study, with genes S6e and S24e found to have multiple splice products in most chordates. In particular for these genes, multiple transcripts (each supported by ESTs) were apparent in many vertebrates, whereas only a single product was predicted in other lineages. In addition, alternative forms of other RPs were found consistently in *X. laevis*, *S. pombe* and *S. cerevisiae*. Low sequence divergence of alternate sequences suggested they were the result of recent gene (genome) duplications, and had not undergone functional diversification (see also Planta and Manger, 1998). Shallow gene duplications are unlikely to have affected phylogenetic analyses at the deep hierarchical levels under investigation here, but comparisons of RP sequences among more closely related taxa would require a robust assessment of orthology using criteria such as locus synteny (although this is not possible with EST data alone). In general, however, our analyses showed that RP genes do not regularly give rise to functional, expressed gene paralogs, and we found no evidence for recent RP gene duplications in Metazoa beyond the few already known. This situation is in contrast to most other types of nuclear genes, where paralogs are commonplace, which has necessitated the removal of numerous genes from other recent analysis of Metazoan relationships (see Wolf et al., 2004; Blair et al., 2005; Philip et al., 2005; Hughes et al., 2006).

Conclusions

The Coelomata versus Ecdysozoa controversy has greatly influenced interpretation of genomic data from model Metazoa. This debate may have been overly polarized regarding the phylogenetic conclusions. The affinity of insects/arthropods with chordates broadly reflects an older view of close relationships of taxa possessing a true body cavity, a coelom, and hence the reference to the "Coelomata hypothesis" in the phylogenomics literature. However, this term is unsatisfactory, as the arrangement of the major animal phyla based on their body cavity was established mostly for practical reasons rather than implied relationships (Hyman, 1951). In fact, the "Coelomata" is not

currently accepted in morphological taxonomy as representing a monophyletic group (e.g., Wägele et al., 1999; Nielsen, 2001). Even the classical morphological schemes of bilaterian phyla (e.g., Nielsen et al., 1996; Fig. 1C) do not recover any clade corresponding to "Coelomata". Given the taxon sampling of most genomic studies, the Ecdysozoa also reflects the hypothesized monophyly of the wider protostome lineage, while support for alternative groupings to the Ecdysozoa are not discussed, such as Articulata that places arthropods with annelids and mollusks (to the exclusion of nematodes and chordates) (see Scholtz, 2002). Worse, the limited taxonomic focus of most phylogenomics studies would not be able to discriminate these traditional schemes from the "new phylogeny", as both imply a closer relationship of arthropods with nematodes than with vertebrates (see Fig. 1C). Hence, the juxtaposition of Coelomata versus Ecdysozoa hypotheses is contrived, and conceived mainly to accommodate an awkward taxon sampling reflecting the limited availability of genomics data.

Yet, the debate about the relationships of major metazoan groups is highly relevant to the broader utility of genome data in phylogenetics, and the difficulty of detecting spurious branch support. Our analyses of a carefully curated subset of RPs show they are suitable phylogenetic markers for resolving basal metazoan relationships, in that they are largely free of paralogy, evolve slowly, and display synergistic phylogenetic signal in combined analyses. The focus of recent studies has been on ever-greater numbers of gene markers and automated data extraction procedures, rather than quality of edits and assessment of congruence in phylogenetic signal. This has overemphasized the power of large numbers of genes and machine edited data without specific tests of the data quality and the success of the algorithms used for data editing. Equally, issues of taxon sampling have been addressed in an intuitive way, driven by data availability, whereby the end-point of the analysis appears to be a topology that conforms to certain notions of the true tree (Philippe et al., 2005). Besides issues of LBA (Brinkmann et al., 2005; Philippe et al., 2005), data exploration will also have to include a careful analysis of emerging phylogenetic support and conflict, as well as paralogy, in a total evidence framework (Figs 4 and 5; Chiu et al., 2006).

Simulations such as those conducted here can play an important part in evaluating whether biases in sequence variation confound the phylogenetic analysis. It remains to be seen if the Ecdysozoa will ever be strongly supported with additional genomic data and, if found consistently with new data, whether this clade reflects phylogenetic history rather than convergences of sequence variation. However, it is clear from the current study that genome-wide biases in sequence evolution are a plausible explanation for the grouping of insects with chordates, rather than nematodes. RP genes are highly expressed and hence likely to be under strong selection to reflect genome specific pressures, while less highly expressed genes may not suffer from this problem to the same degree but instead may be subject to greater levels of paralogy. Only careful investigation of different classes of genes will permit phylogenetic reconstruction of major metazoan clades from genomic data. In addition, most genomic approaches to metazoan systematics still suffer from acute taxon sampling deficiencies. This is expected to have a fundamental effect on phylogenetic estimates, in particular when the distance between the ingroup (bilaterian lineages) and the outgroup (Fungi) is so critical.

Acknowledgments

We thank Kosmas Theodorides, Alessandra de Riva and Monica Mejia-Chang for generating Coleoptera ESTs, Anthony Cognato for assistance with PBS and PHBS analyses, and Wendy Baker (EMBL) for discussions on problems with integrating ESTs and curated sequence databases. We thank Joseph Hughes and Anna Papadopoulou for discussion and help with data generation. We are also grateful to particularly helpful reviewers and their guidance with interpreting recent literature on metazoan relationships. This work was funded by a BBSRC studentship to SJL, and BBSRC grant G14548.

References

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., de Rosa, R., 2000. The new animal phylogeny: reliability and implications. Proc. Natl Acad. Sci. 97, 4453–4456.

Aguinaldo, A.M.A., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489–493.

Baker, R.H., DeSalle, R., 1997. Multiple sources of molecular characteristics and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46, 654–673.

Blair, J.E., Ikeo, K., Gojobori, T., Hedges, S.B., 2002. The evolutionary position of nematodes. BMC Evol. Biol. 2, 1–7.

Blair, J.E., Shah, P., Hedges, S.B., 2005. Evolutionary sequence analysis of complete eukaryotic genomes. BMC Bioinformatics 6, R53.

Brinkmann, H., van der Giezen, M.V., Zhou, Y., Raucourt, G.P.D., Philippe, H., 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54, 743–757.

Brochier, C., Bapteste, E., Moreira, D., Philippe, H., 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18, 1–5.

Chiu, J.C., Egan, L.E.K., M.G., Sarkar, I.N., Coruzzi, G.M., DeSalle, R., 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22, 699–707.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Copley, R.R., Aloy, P., Russell, R.B., Telford, M.J., 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. Evol. Dev. 6, 164–169.

Dopazo, H., Dopazo, J., 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6, Art. no. R41.

Dopazo, H., Santoyo, J., Dopazo, J., 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryotic model species. Bioinformatics 20, 116–121.

Eernisse, D.J., Albert, J.S., Anderson, F.E., 1992. Annelida and arthropoda are not sister taxa – A phylogenetic analysis of spiralian metazoan morphology. Syst. Biol. 41, 305–330.

Felsenstein, J., 1989. PHYLIP: Phylogeneic Inference Package, Version 3.2. Cladistics 5, 164–166.

Foster, P.G., 2004. Modeling compositional heterogeneity. Syst. Biol. 53, 485–495.

Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48, 284–290.

Friedrich, M., Tautz, D., 1997. An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol. Biol. Evol. 14, 644–653.

Gatesy, J., O'Grady, P., Baker, R.H., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15, 271–313.

Giribet, G., 2002. Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? Mol. Phylogenet. Evol. 24, 345–357.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hausdorf, B., 2000. Early evolution of the Bilateria. Syst. Biol. 49, 130–142.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Hughes, J., Longhorn, S.J., Papadopoulou, A., Theodorides, K., de Riva, A., Mejia-Chang, M., Foster, P.G., Vogler, A.P., 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol. Biol. Evol. 23, 268–278.

Hwang, U.W., Kim, W., Tautz, D., Friedrich, M., 1998. Molecular phylogenetics at the Felsenstein zone: approaching the Strepsiptera problem using 5.8s 28S rDNA Sequences. Mol. Phylogenet. Evol. 9, 470–480.

Hyman, L.H., 1951. The Invertebrates, Vol. 2. McGraw-Hill, New York.

Karsi, A., Patterson, A., Feng, J., Liu, Z.Z., 2002. Translational machinery of channel catfish: I. A transcriptomic approach to the analysis of 32, 40S ribosomal protein genes and their expression. Gene 291, 177–186.

Landais, I., Ogliastro, M., Mita, K., Nohata, J., López-Ferber, M., Duonor-Cérutti, M., Fournier, P., Devauchelle, G., 2003. Annotation pattern of ESTs from *Spodoptera frugiperda* cells (Sf9) and analysis of the insect-specific features and unexpectedly low codon usage bias. Bioinformatics 19, 2343–2350.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J., 2000. An optimised protocol for analysis of EST sequences. Nucleic Acids Res. 28, 3657–3665.

Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.

Mallatt, J.M., Garey, J.R., Schultz, J.W., 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phylogenet. Evol. 31, 178–191.

Manuel, M., Kruse, M., Muller, W.E.G., Parco, Y.L., 2000. The comparison of beta-thymosin homologs among Metazoa supports an arthropod–nematode clade. J. Mol. Evol. 51, 378–381.

Meyer, A., Zardoya, R., 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu. Rev. Ecol. Syst. 34, 311–338.

Mushegian, A.R., Garey, J.R., Martin, J., Liu, L.X., 1998. Large-scale taxonomic profiling of eukaryotic model organism: a comparison of orthologous proteins encoded by the human, fly, nematode and yeast genomes. Genome Res. 8, 590–598.

Nielsen, C., 2001. Animal Evolution: Interrelationships of the Living Phyla. Oxford University Press, Oxford.

Nielsen, C., Scharff, N., Eibye-Jacobsen, D., 1996. Cladistic analyses of the animal kingdom. Biol. J. Linnaean Soc. 57, 385–410.

Peterson, K.J., Eernisse, D.J., 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol. Dev. 3, 170–205.

Philip, G.K., Creevey, C.J., McInerney, J.O., 2005. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22, 1175–1184.

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.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

Planta, R.J., Manger, W.H., 1998. The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. Yeast 14, 471–477.

Pruitt, K.D., Katz, K.S., Sicotte, H., Maglott, D.R., 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 16, 44–47.

Quackenbush, J., Cho, J., Lee, D.D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., White, J., 2001. The TIGR gene indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29, 159–164.

Rokas, A., Kruger, D., Carroll, S.B., 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310, 1933–1938.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, J., Baguna, J., Riutort, M., 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc. Natl Acad. Sci. 99, 11246–11251.

Savard, J., Tautz, D., Lercher, M.J., 2006. Genome-wide acceleration of protein evolution in flies (Diptera). BMC Evol. Biol. 6, Art. no. 7.

Schmidt, H.A., Strimmer, K., Vingron, M., Haeseler, A.V., 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analyses and parallel computing. Bioinformatics, 18, 502–504.

Scholtz, G., 2002. The Articulata hypothesis – or what is a segment? Organisms Divers. Evol. 2, 197–215.

Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51, 492–508.

Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics, 17, 1246–1247.

Slater, G., 2000. Algorithms for analysis of ESTs. PhD thesis. University of Cambridge, UK. URL: http://www.ebi.ac.uk/~guy/estate/.

Sorenson, M.D., 1999. Treerot, Version 2.0. Boston University, Boston, MA.

Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis using Parsimony, Version 4.0b. Sinauer Associates, Sunderland, MA.

Telford, M.J., 2004. Animal phylogeny: Back to the Coelomata? Curr. Biol. 14, 274–276.
Templeton, A.R., 1983. Phylogenetic inference from restriction site endonuclease cleavage site maps with particular reference to humans and apes. Evolution 37, 221–244.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
Uechi, T., Tanaka, T., Kenmochi, N., 2001. A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. Genomics 72, 223–230.
Wägele, J.W., Erikson, T., Lockhart, P., Mishof, B., 1999. The Ecdysozoa: artifact or monophylum? J. Zool. Syst. Evol. Res. 37, 211–223.
Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.
Wolf, Y.I., Rogozin, I.B., Koonin, E.V., 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14, 29–36.
Zhang, Z., Harrison, P., Gerstein, M., 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human gene. Genome Res. 12, 1466–1482.
Zrzavy, J., 2001. Ecdysozoa versus Articulata: clades, artifacts, prejudices. J. Zool. Syst. Evol. Res. 39, 159–163.

Supplementary material

The authors have provided the following supplementary material, which is available alongside the article at http://www.blackwell-synergy.com.

Appendix S1 (a) table of accession details and curations, and (b) table of branch support at nodes not shown in Fig. 3.

Figure S1. Phylogenetic trees of 25 RP genes in separate analysis using parsimony [with exact Coleoptera given in the supplementary information; where Poly (Cu.) is Polyphaga; Cucujiformia, and (El.) is Elateriformia]. For each gene, the number of equally parsimonious solutions is given below each subheader (I–XXV). Values above nodes are branch (Bremer) support, below nodes are bootstrap scores.