Near Intron Pairs and the Metazoan Tree

Jörg Lehmann[a], Peter F. Stadler[a,b,c,d,e,f], Veiko Krauss[a,*]

[a]*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany*
[b]*Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany*
[c]*RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, 04103 Leipzig, Germany*
[d]*Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria*
[e]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark*
[f]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Gene structure data can substantially advance our understanding of metazoan evolution and deliver an independent approach to resolve conflicts among existing hypotheses. Here, we used changes of spliceosomal intron positions as novel phylogenetic marker to reconstruct the animal tree. This kind of data is inferred from orthologous genes containing mutually exclusive introns at pairs of sequence positions in close proximity, so-called near intron pairs (NIPs). NIP data were collected for 48 species and utilized as binary genome-level characters in maximum parsimony (MP) analyses to reconstruct deep metazoan phylogeny. All groupings that were obtained with more than 80% bootstrap support are consistent with currently supported phylogenetic hypotheses. This includes monophyletic Chordata, Vertebrata, Nematoda, Platyhelminthes and Trochozoa. Several other clades such as Deuterostomia, Protostomia, Arthropoda, Ecdysozoa, Spiralia, and Eumetazoa, however, failed to be recovered due to a few problematic taxa such as the mite *Ixodes* and the warty comb jelly *Mnemiopsis*. The corresponding unexpected branchings can be explained by the paucity of synapomorphic changes of intron positions shared between some genomes, by the sensitivity of MP analyses to long-branch attraction (LBA), and by the very unequal evolutionary rates of intron loss and intron gain during evolution of the different subclades of metazoans. In addition, we obtained an assemblage of Cnidaria, Porifera, and Placozoa as sister group of Bilateria + Ctenophora with medium support, a disputable, but remarkable result. We conclude that NIPs can be used as phylogenetic characters also within a broader phylogenetic context, given that they have emerged regularly during evolution irrespective of the large variation of intron density across metazoan genomes.

Keywords: intron evolution, molecular phylogenetics, maximum parsimony, metazoan phylogeny, Ecdysozoa, rare genomic changes

1. Introduction

The evolutionary relationships of metazoan phyla still constitute a challenge for both morphological and molecular-based analyses. The traditional view arranges bilaterian metazoans into acoelomates, pseudocoelomates, and coelomates. Starting with the work of Aguinaldo et al. (1997), sequence data, initially mostly rDNA, have been used to establish a "new animal phylogeny" with far-reaching consequences: (1) the protostomes were divided into ecdysozoans (Aguinaldo et al., 1997) and lophotrochozoans (Halanych et al., 1995), and (2) several phyla representing apparently lower grades of complexity (Platyhelminthes, Nemertea, and Nematoda) were relocated amongst the coelomate groups at the crown of the tree. In contrast, some studies employing genomic datasets containing only a few taxa (e.g. Wolf et al., 2004) supported the monophyly of coelomates. Later studies, however, have shown that these results are likely artefacts, misled by a faster evolution of some genomes, such as that of *Caenorhabditis elegans* (Philippe et al., 2005). These are thus often excluded from phylogenetic datasets. In addition, conflicting signals are often obtained from mitochondrial, nuclear rRNA, phylogenomic and also morphological data (Trautwein et al., 2012). Despite a plethora of studies based on both molecular and morphological data, a consensus on the phylogenetic tree of metazoan phyla is still not in sight (Edgecombe et al., 2011). This concerns in particular the non-bilaterians (Dunn et al., 2008; Schierwater et al., 2009; Pick et al., 2010) and the Lophotrochozoa (Hejnol, 2010).

Characters resulting from structural changes of the genomic sequence, so-called rare genomic changes (RGCs), such as coding insertions/deletions (indels) (Belinky et al., 2010), spliceosomal intron positions (Irimia and Roy, 2008), and positions of mobile genetic elements (Kriegs et al., 2006, 2010), are expected to be less prone to homoplasy than substitution patterns of sequence data and hence provide valuable additional information to resolve conflicts in phylogenetic tree reconstruction. For holometabolic insects, novel phylogenetic hypotheses have been introduced on the basis of such characters, for example the basal position of Hymenoptera (Krauss et al., 2005). Later,

*Corresponding author
Email address: krauss@rz.uni-leipzig.de (Veiko Krauss)

this proposal received strong support by sequence-based analyses of single-copy nuclear genes and additional intron position data (Savard et al., 2006; Zdobnov and Bork, 2007; Krauss et al., 2008; Wiegmann et al., 2009). Another study used retrotransposon insertions to improve our knowledge about the basal branching order of rodents (Churakov et al., 2010). Earlier attempts to reconstruct the radiation of rodents are well-known to have suffered from long-branch attraction (LBA) artefacts.

The present study utilizes the conservation of positions of spliceosomal introns among orthologous coding sequences (CDS). Intron positions have already been used by several authors to resolve problematic branches of the metazoan tree (for review see Irimia and Roy, 2008). For instance, an intense debate emerged about the concepts of Ecdysozoa (Roy and Gilbert, 2005) and Coelomata (Zheng et al., 2007) using intron position data. Roy and Gilbert (2005) supported the taxon Ecdysozoa using a pattern of intron conservation. This was criticized by Zheng et al. (2007) by showing that intron loss rates within specific branches are strongly correlated. These authors argued that high rates of independent intron losses within the used nematode and arthropod species had misled the former study. However, in turn, Roy and Irimia (2008) identified several weaknesses of the latter analysis, among them biases in the procedure used to differentiate between intron gain and loss. Pointing to both large intron loss and gain rate variations, Roy and Irimia (2008) avoided a clearcut conclusion about the Ecdysozoa/Coelomata problem.

In order to reduce the impact of homoplastic characters due to parallel intron gains or losses, we specifically consider pairs of nearby introns. More precisely, a near intron pair (NIP) consists of two intron positions in an alignment of two or more orthologous genes that are separated by a small number of nucleotides. Exons smaller than about 50 nt are relatively rare (Saeys et al., 2007) and in general functionally detrimental (Weir et al., 2006). The two nearby intron positions are thus very unlikely to have coexisted. Under the assumption that parallel intron gain is very rare, a NIP can be used to parsimoniously infer an edge of the phylogenetic tree along which both intron loss and gain must have occurred, separating the species sharing one of the positions from those that share the other.

In previous work, we found some evidence that NIPs arise not only from uncoupled, successive processes of intron loss and intron gain, but also from intron sliding (Krauss et al., 2005, 2008; Lehmann et al., 2010). For *Drosophila* we could show that some of the younger NIPs were indeed caused by shifts of splice donor and acceptor sites in relation to conserved CDS (Lehmann et al., 2010). In the same study, we used NIPs for a systematic investigation of intron gain mechanisms in *Drosophila* .

Encouraged by the successful application of NIPs to the phylogeny of holometabolan insects (Krauss et al., 2008; Niehuis et al., 2012), we here try to resolve the phylogenetic tree of animals based exclusively on NIP data from 45 metazoan and 3 outgroup taxa. Our results demonstrate the usefulness of NIPs as phylogenetic marker also for deep metazoan phylogeny. In particular, we evaluate the Ecdysozoa hypothesis, as well as the general agreement of our tree reconstructions with current proposals of metazoan phylogeny.

2. Material and methods

2.1. Compilation of Ortholog Data Set

Initially, we retrieved orthologous protein-coding genes from the Ensembl Compara database (release 67, May 2012) (Flicek et al., 2011) in the following manner: For a set of 8 selected query species (*Acyrthosiphon pisum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Ixodes scapularis*, *Nematostella vectensis*, *Schistosoma mansoni*, *Strongylocentrotus purpuratus*, and *Trichoplax adhaerens*), all protein-coding gene IDs with the status 'Known' were determined using Ensembl Biomart. Then, these reference genes and their predicted 1:1 orthologs within the Ensembl Metazoa (v14) and Ensembl Core (v67) databases were retrieved from the following 29 taxa: *Acyrthosiphon pisum*, *Aedes aegypti*, *Amphimedon queenslandica*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori*, *Caenorhabditis brenneri*, *Caenorhabditis briggsae*, *Caenorhabditis elegans*, *Caenorhabditis japonica*, *Caenorhabditis remanei*, *Ciona intestinalis*, *Ciona savignyi*, *Culex quinquefasciatus*, *Danio rerio*, *Daphnia pulex*, *Drosophila melanogaster*, *Homo sapiens*, *Ixodes scapularis*, *Monodelphis domestica*, *Nematostella vectensis*, *Pediculus humanus*, *Pristionchus pacificus*, *Schistosoma mansoni*, *Strongylocentrotus purpuratus*, *Takifugu rubripes*, *Tribolium castaneum*, *Trichinella spiralis*, and *Trichoplax adhaerens*. For each gene, only the transcript coding for the longest isoform was selected. If a gene was contained in more than one of these putative ortholog groups, all the affected groups were excluded from the dataset to avoid the inclusion of paralogs. Finally, only the 4,405 ortholog groups containing genes from at least 50% of the species were retained for further processing in order to limit the amount of missing data. In order to extend this core dataset, we followed two different approaches, depending on the target species:

A targeted search for orthologs based on available gene builds was performed for the 13 additional species *Branchiostoma floridae*, *Brugia malayi*, *Capitella teleta*, *Coprinopsis cinerea*, *Dictyostelium purpureum*, *Helobdella robusta*, *Heterorhabditis bacteriophora*, *Lottia gigantea*, *Meloidogyne hapla*, *Meloidogyne incognita*, *Monosiga brevicollis*, *Nasonia vitripennis*, and *Schistosoma japonicum* (see Supplementary Tables S1–S2). For this purpose, we used the `hamstrsearch_local` package (Ebersberger et al., 2009) (`HaMStR v8b`) to determine reliable ortholog additions. The pipeline generates profile hidden Markov models (HMMs) from the ortholog groups of the core dataset and uses these to retrieve candidate hits within each target proteome using `hmmsearch` (`HMMER 3.0` package, `http://hmmer.org`). These are automatically checked for reciprocity using `blastp` (Camacho et al., 2009) against a subset of reference species. Here, the reference species more closely related to the target species was used preferentially (see Supplementary Table S2 for details). Target proteins that could not be uniquely assigned to a single query protein were excluded from the resulting dataset.

For the remaining 6 target species (*Aplysia californica*, *Heterorhabditis bacteriophora*, *Mnemiopsis leidyi*, *Rhodnius prolixus*, *Saccoglossus kowalevskii*, and *Schmidtea mediterranea*), no gene builds were available, hence we used BLAST (Camacho et al., 2009) to identify orthologs. For this purpose, we performed `tblastn` searches for each target species using the full proteome sets of 4 different reference species from the core dataset (see Supplementary Tables S1, S3) to ensure that each ortholog group is represented by at least one query. For a particular ortholog group, the query sequence that was actually used for the `tblastn` retrieval was selected based upon availability and a ranked order of the four query species, such that the more closely related query species was selected preferentially. Best-hit genomic target regions were automatically extracted, and CDS were predicted with `exonerate` (Slater and Birney, 2005) (spliced-alignment using the `protein2genome` model) based on the query protein. Occasional frame shifts due to insertions or deletions were compensated by short artificial gaps in the CDS sequence and annotation. A refinement of CDS predictions was obtained for a subset of candidates for which appropriate target proteins and/or mRNAs were available from NCBI databases. In this case the spliced-alignment with `exonerate` was based on the protein of the target species rather than the homologous query protein, see Supplementary Table S3. Each CDS prediction within these 6 species was required to

1. have no overlap with a prediction from another query protein with a larger `tblastn` bit score,
2. have a query coverage of at least 50%,
3. have a mean identity of at least 25% (as measured by the `tblastn` HSPs)

In case of multiple `tblastn` predictions for a target species, only the best-scoring one per ortholog group was retained. Each CDS prediction was checked by reciprocal `blastp` of its translation to the query proteome, and only retained if the initial query was returned as best hit. To test whether the usage of `blastp` for these 6 species has an impact on the resulting NIP phylogeny, we used `HaMStR` to identify putative paralogs here. Removal of these sequences, however, resulted only in a small reduction of the informative NIP dataset (0.2%) and did not affect the topology of the NIP-inferred tree.

2.2. Translated Alignments and Intron Position Mapping

CDS were compiled for each transcript using the available CDS annotation. Similarly to the procedure outlined in Lehmann et al. (2010), the `transAlign` program (Bininda-Emonds, 2005) was used to construct a codon-based multiple alignment for each of the 4,405 ortholog groups. This tool translates nucleic acid sequences to peptide sequences, invokes `Muscle` (Edgar, 2004) to generate a protein alignment and then back-translates them to the corresponding CDS alignment. In addition to the protein alignment with `Muscle`, the realignment tool of Csurös et al. (2007) was employed. It allows to rescore an existing amino acid alignment with intron positions annotated for each individual sequence while at the same time attempting to align the positions of introns whenever possible. We used the standard parameters, which give only a small bonus to aligned intron positions. Any CDS predictions containing internal stop codons were excluded from the translated alignments.

Intron positions were mapped to the codon-based sequence alignments and labeled with the codon position of a reference sequence as defined within each alignment (preferably from *D. melanogaster*, *C. elegans*, or *H. sapiens*). From each alignment, all intervals were extracted that contained at least two intron positions separated by at most 70 alignment columns. We called such alignment intervals near intron pair regions (NIP regions). Each alignment of such a NIP region includes 30 nt flanking CDS alignment sequence around the outmost intron positions. Sequences consisting of gaps only were excluded from these regions.

2.3. Extraction and Selection of NIP Characters

Finally, NIP characters were extracted from NIP regions by collecting each pair of intron positions that fulfilled the distance constraint of <32 nt (Lehmann et al., 2010) for all sequences of a region (i.e., excluding gap characters). NIP characters having both introns present in one of the sequences were removed entirely from the data set. Each NIP character was additionally checked for the local quality of the amino acid alignment around both intron positions. Taxa that are not part of the NIP were not considered in this context of alignment quality. The local quality of the sub-alignment was determined as the average relative sum-of-pair score (see paragraph "Scoring of Protein Alignments") for a sequence window of 3 amino acids around the intron positions, similar to the method employed by Wilkerson et al. (2009). The average-score threshold was set to 0.5. Furthermore, we required the sub-alignment to be gap-free within this sequence window.

A NIP character is encoded by a column in the data matrix containing symbol '1' for species having the upstream intron, symbol '2' for species having the downstream intron of the pair, and a '?' (missing data symbol) for species not contributing with an intron (optionally encoded as state '0') or not part of the NIP region alignment.

2.4. Tree Searches and Testing

Maximum parsimony (MP) tree searches and bootstrap analyses based on NIP character matrices (character type: unordered, i.e. equal state transition costs, Wagner parsimony) were performed using PAUP* 4.0b10 (Swofford, 2003). Default settings for the heuristic search were random stepwise addition and 1,000 replicates with the tree-bisection-reconnection (TBR) branch-swapping option. For bootstrapping 1,000 replicates and simple stepwise addition were used in combination with TBR. The same settings for tree search and bootstrapping were employed for the additional intron presence/absence data analyses using both Dollo parsimony and Wagner parsimony (Supplementary Figures S8–S9). The (ensemble) consistency index (CI) and the retention index (RI) were determined with PAUP* considering only the parsimony-informative characters. Recall that CI measures the level of homoplasy while RI measures the amount of synapomorphy (Farris, 1989). Contradictory hypotheses were evaluated by comparing the total

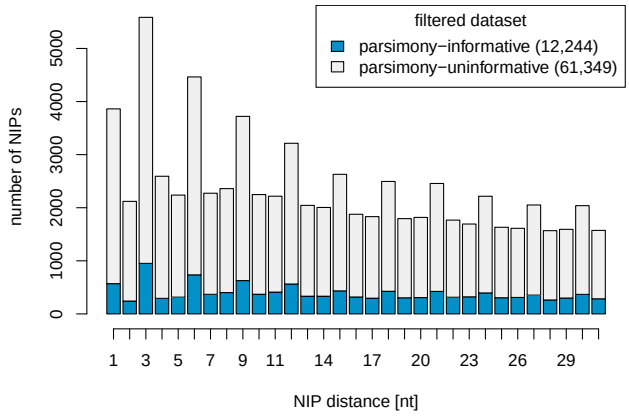

Figure 1: Intron distance distribution of the NIP dataset. The parsimony-informative NIP subset (blue) and the remaining NIPs are grouped according to intron position distances in nucleotides.

tree lengths of the MP topologies with those of constrained MP topologies. Statistical significance was assessed by the Templeton test (Templeton, 1983) and the winning-sites test (Prager and Wilson, 1988) as implemented in PAUP*.

Sequence-based phylogeny reconstructions were conducted with `RAxML` v. 7.3.2 (Stamatakis, 2006) using the WAG substitution model as selected with `ProteinModelSelection.pl` (rapid bootstrapping option, 100 replicates), and `PhyloBayes` (Lartillot et al., 2009) (v. 3.2e; with default settings; convergence of two runs were assumed at a 'maxdiff' value of 0.12, a 'maximum discrepancy' <0.3, a 'minimum effective size' >50. The consensus tree was built from ca. 36,000 sampled trees). For the sequence-based analyses, data sets were selected for almost complete taxon coverage (>41 taxa), and the amino acid alignment positions were trimmed with `Gblocks` (Talavera and Castresana, 2007) using default settings (b3=8, b4=10).

2.5. Scoring of Protein Alignments

For scoring of alignment quality, we calculated for each site of the (partial) multiple amino acid sequence alignment a relative sum-of-pair score using the BLOSUM45 substitution matrix of Clustal W (Larkin et al., 2007) (modified for positive pair-scores) and weighting with the maximally possible sum-of-pair score based on the most abundant amino acid(s). The final average score for the complete alignment was obtained by averaging the relative sum-of-pair score over all sites.

3. Results

3.1. Collection and Characterization of a Large NIP Data Set

Starting from selected Ensembl Compara ortholog predictions, we compiled a set of orthologs covering 48 taxa, comprising 12 metazoan phyla: Cnidaria, Placozoa, Ctenophora, Porifera, Annelida, Mollusca, Platyhelminthes, Chordata, Echinodermata, Hemichordata, Arthropoda, and Nematoda (see Supplementary Table S1, Supplementary Material online). *Monosiga brevicollis* (Choanoflagellata), *Coprinopsis cinerea* (Fungi), and *Dictyostelium purpureum* (Amoebozoa) were added as outgroups.

The automated alignment and NIP extraction pipeline (see Materials and Methods) produced 49,129 partial CDS alignment regions that contain potential near intron pairs (NIP regions, see Supplementary Material 1, Supplementary Material online, for a subset of these regions). Consistent with our previous study on *Drosophila* (Lehmann et al., 2010), the fraction of NIP candidates that need to be excluded from the analysis due to short exons, i.e., because the alignment contains a taxon in which introns are present at both positions, increases for NIP distances of more than some 30 nt (Supplementary Figure S2). Thus, in contrast to Krauss et al. (2008) but in concordance with Lehmann et al. (2010), we here only consider NIPs of distance <32 nt, to further limit the amount of homoplasy in the data set.

Intriguingly, the dataset of used NIPs reveals that distances of 1 nt and of multiples of 3 nt are more abundant than others, especially for very short distances (Figure 1). This may be a consequence of the unequal distribution of intron phases (0:1:2 = 51:27:22) already described by Long and Deutsch (1999) using an early eukaryotic splicing database. The inequalities of this distribution were shown to be due to biased intron gain (Qiu et al., 2004; Nguyen et al., 2006). We evaluated the phase distribution of all introns from our initial ortholog data set and obtained a ratio similar to data of Qiu et al. (2004) (0:1:2 = 50:26:24). Thus, the observed general excess of NIP distances of multiples of 3 nt can be ascribed mainly to the increased probability that two adjacent intron positions have the same phase, which is for our data set 0.38 (distance is multiple of 3, i.e. phase difference is 0), in contrast to 0.31 for phase differences of 1 or 2, respectively.

The additional excess for NIP distances of 1 nt and lower multiples of three is consistent with previous studies (Rogozin et al., 2000; Krauss et al., 2008; Lehmann et al., 2010), which provided evidence for intron sliding, a possible mechanism for the emergence of NIPs that leads to a preference of these distances. In coding regions there is a very strong selection pressure to preserve the reading frame, hence the shift from a splice site to another one located a multiple of three nucleotides away is favoured, allowing the independent migration of donor and acceptor sites. Intron sliding thus could explain the remaining NIP number differences between the distances.

By extracting all possible pairs of mapped intron positions passing the alignment quality filter (see Materials and Methods), we retained 76,150 NIPs. For 2,557 (3.4%) of these, both introns were present in one or more taxa. Most of these short internal exons (<32 nt) were found in *Coprinopsis cinerea*, *Heterorhabditis bacteriophora*, *Trichoplax adhaerens*, and *Ciona intestinalis*, participating in more than 6% of the cases, respectively. All these NIPs were excluded from the subsequent phylogenetic analysis. We finally arrived at a dataset of 73,593 NIPs, of which 12,244 are parsimony-informative. Figure 2 displays an example NIP region, containing a NIP in support of Ecdysozoa.

Figure 2: Example NIP region. The parsimony-informative NIP character `Smp_150040.Mbre.Mbre.300-0_307-2` supports that the last common ancestor of arthropods and nematodes (i.e. of Ecdysozoa) is not an ancestor of Deuterostomia. Intron positions are indicated by lowercase nucleotides. Sequence IDs were replaced by full taxon names, and only the section relevant for the selected NIP is shown.

3.2. Parsimony Search

We conducted MP tree searches and bootstrap runs based on 12,244 parsimony-informative NIPs and obtained the strict consensus MP tree shown in Figure 3.

To compare our NIP-based tree to contemporary sequence-based analyses, we mapped our data to a tree topology that combines the currently preferred metazoan relationships as proposed by Lartillot and Philippe (2008), Hejnol et al. (2009), Mortazavi et al. (2010), and Meusemann et al. (2010) (Figure 4). Here, the tree length is 13,968, i.e. 130 steps longer than the unconstrained MP trees (strict consensus shown in Figure 3). According to Templeton and Winning-site tests, this difference is significant (Table 1).

3.3. Extended Parsimony Approach

A potential shortcoming of the reported analysis is that simple intron losses within the tree cannot be used as phylogenetic information. To overcome this limitation, we included the absence of both intron positions of a NIP as a third character state (zero) in an extended parsimony approach, instead of accounting for it as missing data. However, the standard Dollo-Parsimony approach as implemented in PAUP* (often applied to such presence/absence data) is unsuitable for such a character coding, since regain of the same intron position is very unlikely and cannot be penalized, given two states for the presence of introns.

Hence, we adapted Sankoff's parsimony algorithm (Sankoff, 1974; Sankoff and Rousseau, 1975) for scoring state transitions for each character. Here, changes to and from the zero state have a cost of one, whereas changes directly between the two intron states have a cost of two. Each intron state is allowed to be introduced (along the tree edges) only once without additional costs. Whenever a change to the same intron state occurs more than once in the tree, a large additional penalty is added that explicitly scores the homoplasy. We performed three different heuristic runs using the full NIP dataset (73,593 NIPs) with penalties 1000, 100, and 10, respectively. To obtain bootstrap values was not possible due to the large runtime for one run (about 170h). The results of this more inclusive approach, therefore, are of limited value (Supplementary Figure S3–S5). All three runs resulted in a significantly worse topology than found using (unordered/Wagner) two-state parsimony (Figure 3). These results suggest that the incorporation of absence states does not improve tree inference.

3.4. Sequence-based Phylogeny for Comparison

In order to compare the NIP-based results with classical sequence-based results, we conducted maximum likelihood (ML) and Bayesian (BI) analyses for tree reconstruction based on the amino acid alignments from the same ortholog dataset. For this purpose, we filtered the 4,405 alignments to contain at least 42 taxa. This resulted in 191 alignments. After trimming alignment positions using `Gblocks`, the concatenated dataset contained 9,121 amino acid alignment sites from 134 genes. Figure 5 displays the `RAxML` bootstrap consensus tree obtained from ML tree searches. The resulting topology (as well as that

Issue	Constraint	MP steps	MP trees	Templeton	Winning-sites
	Fixed topology from Figure 4	13,968	1	<0.0001***	<0.0001***
1a	((*Mnemiopsis*, *Amphimedon*, *Trichoplax*, Cnidaria))	13,873	5	0.0001***	0.0002***
1b	((*Amphimedon*, *Trichoplax*, Cnidaria, Bilateria))	13,856	10	0.1939	0.2445
1c	((*Mnemiopsis*, *Trichoplax*, Cnidaria, Bilateria))	13,842	5	0.2850	0.4240
1d	((*Mnemiopsis*, *Amphimedon*, Cnidaria, Bilateria))	13,851	5	0.0016**	0.0023**
1e	((*Mnemiopsis*, *Trichoplax*, *Amphimedon*, Bilateria))	13,850	10	0.0105*	0.0192*
1f	((Cnidaria, Bilateria))	13,881	10	0.0001***	0.0002***
2	Deuterostomia	13,846	10	0.0455*	0.0768
3a	Coelomata: ((Arthropoda, Trochozoa, Deuterostomia))	13,961	5	<0.0001***	<0.0001***
3b	Ecdysozoa: ((Arthropoda, Nematoda))	13,858	5	0.0168*	0.0232*
4a	Arthropoda	13,893	5	<0.0001***	<0.0001***
4b	Holometabola	13,848	10	0.2809	0.3318
4c	((*Aedes*, *Culex*))	13,839	5	0.5637	1.0000
5a	((*Meloidogyne*, *Pristionchus*, *Heterorhabditis*, *Caenorhabditis*))	13,842	5	0.4142	0.5413
5b	((*C. brenneri*, *C. remanei*, *C. briggsae*))	13,840	5	0.5271	0.7539
6	Spiralia	13,849	14	0.1790	0.2218

Table 1: Comparison of constrained MP topologies with the unconstrained MP topologies using the NIP dataset (12,244 informative NIPs). Levels of significance in the Templeton (Templeton, 1983) and Winning-sites (Prager and Wilson, 1988) tests in comparison to the unconstrained MP topologies (5 trees with 13,838 steps each, strict consensus shown in fig. 3) are indicated by stars. The table displays only the largest *p*-value observed among all pairwise comparisons of unconstrained and constrained MP topologies, respectively.

obtained from the BI analysis, see Supplementary Figure S6) is similar to that taken from the literature (Figure 4), except with respect to the positions of *Mnemiopsis*, *Trichoplax*, *Brugia*, and the Ambulacraria (*Saccoglossus* and *Strongylocentrotus*).

4. Discussion

Here, we extracted for the first time near intron pairs (NIPs) from a broad collection of genomes and used them as binary genome-level character in maximum parsimony analyses to explore the information value of NIPs to reconstruct deep metazoan phylogeny. The resulting tree deviates remarkably from contemporary hypotheses of metazoan relationships. Notably, the unusually distributed taxa *Mnemiopsis*, *Ixodes*, and Ambulacraria prevented the tree from resolving major clades such as Bilateria, Ecdysozoa, Arthropoda, and Deuterostomia. In addition, we obtained an assemblage of Cnidaria, Porifera, and Placozoa as sister group to Bilateria + *Mnemiopsis*, which is a disputable, but remarkable result. We conclude that NIPs can in principle be used as phylogenetic characters within a broader phylogenetic context as the corresponding changes of spliceosomal intron positions seem to have happened more or less regularly during metazoan evolution irrespective of the large variation of intron density across genomes.

The taxa in this analysis were selected to include mainly deep metazoan branches, e.g. we included only a few vertebrates and only one representative of *Drosophila* (for a NIP-based phylogeny of the genus *Drosophila* see Lehmann et al. (2010)). On the other hand, within some important lineages the number of available genomes is still very limited (e.g. Spiralia) and species sampling in part concentrates on uncommon representatives such as *Schistosoma*. Furthermore, the coverage of orthologous genes in our dataset varies substantially between species (e.g. orthologs for *A. californica* are only present in a fifth of the dataset). In addition, the intron densities observed within these ortholog predictions vary considerably (e.g. from less than 3 introns per gene for *S. mediterranea*, *D. purpureum*, and *M. leidyi*, up to 11 or 13 introns per gene for *T. rubripes* and *P. pacificus*), see Supplementary Figure S7 for a comparison. In general, large differences of intron densities as expected between more distantly related species may pose a problem for the correct inference of deep relationships using intron positions as markers (Rogozin et al., 2005). Specifically, we observed within our dataset of 12,244 parsimony-informative NIP characters a rate of 78.9% missing data. Slightly more than half of these cases can be attributed to the absence of introns (43.2%), the remaining 35.7% result from the absence of orthologous sequences.

Another caveat for our analysis is that NIP characters in part depend on each other. This may be the case when NIP regions contain more than one NIP. Among the 12,244 parsimony-informative NIPs, we found 3,445 intron positions that are used more than once: 3,049 positions appear in 2 NIPs, 352 are used three times, 33 four times, 7 five times, and 4 are used six times. Our dataset thus comprises only 20,588 intron positions instead of the theoretically expected 24,488.

Although it is possible in principle to encode groups of overlapping NIP characters as a single multi-state character to enforce character independence, this appears to be impractical. The reason is that there is a large number of different local situations that give rise to many different arrangements of possible transitions between the multiple character states. Alternatively, characters may be reduced to the subset where each intron is used only once. However, applied to our dataset, this reduced the amount of phylogenetic information by more than one third and the topology of the resulting tree was further impaired (see Supplementary Figure S10).

Partial dependency of characters does not appear to be a dramatic problem in general: nucleotide and protein sequence data cannot be expected to be free of correlations between adjacent characters either. The variability of a site within a protein or an RNA sequence depends upon its functional and structural context (Savill et al., 2001; Conant and Stadler, 2009) and hence to a certain extent on its neighbors. In the context of proteins, this

is the biochemical foundation of the covarion model of molecular evolution (Penny et al., 2001). In the case of RNA, where the dependence of base paired nucleotides is nearly complete, specialized substitution models can be used (Jow et al., 2002). These still neglect the weaker correlations between adjacent positions resulting from base pair stacking.

Irrespective of these complications, our NIP-based MP analysis (Figure 3) suggests monophyletic, well-established clades such as Cnidaria, Ambulacraria, Chordata, Vertebrata, Pancrustacea, Nematoda, Platyhelminthes, and Trochozoa. However, only the bootstrap support of Ambulacraria, Vertebrata, Pancrustacea, and Platyhelminthes, respectively, is above 90%. We found that all branchings of our strict consensus tree with more than 80% bootstrap support (Figure 3) are consistent with currently published phylogenies (Figure 4). Some other branches diverge from this super tree. To evaluate the compatibility of our analysis results with current phylogenetic hypotheses, we conducted constrained MP analyses for selected groupings and compared the results with the unconstrained topologies (Table 1). We discuss them in the background of current metazoan tree reconstructions:

1. According to the NIP-based MP analysis, the cnidarians *Nematostella vectensis* and *Hydra magnipapillata*, the placozoan *Trichoplax adhaerens*, and the sponge *Amphimedon queenslandica* are grouped together as sister clade to all other animals in our trees with a bootstrap support of more than 77%. Monophyletic diploplasts would be in agreement with a previous analysis that postulates an early separation of diploplastic animals from a bilaterian ancestor (Schierwater et al., 2009), however, in our analysis the diploblast *Mnemiopsis* clearly is misplaced as a supposed sister of the Pancrustacea. In contrast, Mallatt et al. (2010) support a sister relationship between *Trichoplax* and Cnidaria as well as between Porifera (as represented by *Amphimedon*) and all other metazoans. Other sequence-based phylogenetic analyses (Srivastava et al., 2010; Pick et al., 2010) propose, instead, the placement of either *Trichoplax* or Cnidaria as sister to the group of all other Eumetazoa, respectively, and Porifera as earliest branching metazoan lineage. Philippe et al. (2011) question the results of Schierwater et al. (2009) on the grounds of several methodological issues that may have resulted in a strong non-phylogenetic signal due to scarce taxon sampling and a weak phylogenetic signal as a consequence of short internal branches.

Constrained tree searches using NIP data (Table 1) with *Mnemiopsis* (1b) or *Amphimedon* (1c) as most basal group of metazoans could not reject these constrained tree variants as alternatives to the unconstrained MP tree topologies. Only groupings with *Trichoplax* (1d) or Cnidaria (1e) as most basal group of metazoans as well as monophyletic diploblasts including *Mnemiopsis* seem to require significantly longer trees, respectively, but results have to be considered with caution due to small number of cases with differences. Similarly inconclusive are the results of our sequence-based analyses. `PhyloBayes` placed *Amphimedon* and *Mnemiopsis* (unresolved) at the base of the animal tree, followed by *Trichoplax* and then Cnidaria resulting from basal splits, respectively, whereas `RAxML` obtained an implausible distribution of these species with low bootstrap support (Figure 5 and Supplementary Figure S6).

The number and distribution of taxa available for early diverging metazoan lineages is still insufficient for a convincing analysis. More problematically, all five diploblastic taxa exhibit much fewer shared intron gains (on average 0.3% of all parsimony-informative NIPs) compared to the other metazoan genomes (on average 5.5% of all parsimony-informative NIPs, Figure 6), resulting in a much weaker phylogenetic signal than available elsewhere in the metazoan tree. Maximum parsimony is well-known for being very sensitive to LBA (Felsenstein, 1978). The very low abundance of phylogenetically informative, novel introns in *Mnemiopsis* combined with some parallel intron gains here and in some pancrustaceans could have cause the misplacement of Ctenophores. Rare cases of detected novel introns might be caused by frequent, independent intron losses within all clades. This necessarily causes a shortage of traceable shared intron gains in early diverging branches. Thus, the inclusion of additional taxa diverging from basal splits may help to resolve the positions of diploblastic animals in particular in a NIP-based tree.

2. In the MP tree (Figure 3), *Strongylocentrotus purpuratus* (Echinodermata) and *Saccoglossus kowalevskii* (Hemichordata) are not grouped as sister to the remaining deuterostomes, but as sister to all remaining bilaterian taxa + *Mnemiopsis*, thus contradicting a monophyletic clade of deuterostomes. A similar misplacement of these taxa was found by Nesnidal et al. (2010), here seemingly reflecting a compositional bias in amino acid composition. Both species show much fewer intron gains than all other deuterostomian taxa in our tree (Figure 6). Thus, the misplacement in the NIP dataset result might be due to high loss rates of conserved introns combined with a very limited gain of new introns during early deuterostomian evolution. Indeed, the difference to the constrained tree is not significant (Table 1). Interestingly, urochordates (*Ciona*), a taxon which was often misplaced in sequence-based phylogenies (e.g. Bourlat et al., 2008; Mallatt et al., 2010), is consistently found within the chordate partitions of the trees (Figure 3), in agreement with the new chordate phylogeny (Delsuc et al., 2006). Here, intron evolution in the inferred common ancestor of vertebrates and *Ciona* provided sufficient synapomorphic intron position changes to resolve this branch.

3. Enforcing a Coelomata constraint (3a, Table 1) is significantly worse compared to the unconstrained MP trees (tree lengths 13,961 *vs.* 13,838, $P < 0.0001$). In contrast, an Ecdysozoa constraint results in a tree which is only 20 steps longer than the unconstrained one, and this difference is less significant. Thus, NIP data prefer the more recent morphological concept of moulting animals against the Coelomata (see Mallatt et al., 2010, and references therein).

4. Some arthropod species show unusual positions in the MP tree (Figure 3). First, the mite *I. scapularis* is placed as sister to all other ecdysozoans + Platyhelminthes, instead of at the basal split from all other arthropods. A tree search enforcing

Arthropoda yields significantly longer trees (4a, Table 1). The problematic position of *I. scapularis* is likely caused by the unusually small fraction of younger introns present, compared to all other ecdysozoan species analyzed (Figure 6).

Second, the monophyly of Hexapoda and of holometabolan insects is not recovered, by grouping the hymenopterans together with paraneopterans and *Daphnia* as sister to all other holometabolans. Low support values and insignificantly longer trees when enforcing Holometabola (4b, Table 1) point to a currently unknown, but specific problem, as synapomorphic introns were abundantly found in the relevant genomes.

Third, the mosquito species *A. aegypti* and *C. quinquefasciatus* failed to group together, but also do not require longer trees in case of a constraint (4c). Difficulties to arrange the three mosquito species as expected might be due to the relatively short evolutionary times between them in contrast to the relatively large distance to the next-related species *D. melanogaster*. As intron loss and gain have occurred in very unequal rates during evolution, NIP characters are supposed to be especially sensitive to such distance effects.

5. A large number of intron gains supports the generally well resolved branches of nematode genera (Figures 3 and 6), in accordance with the high speed of intron evolution in this group (Coghlan and Wolfe, 2004; Cho et al., 2004). The branching order of *Brugia malayi* and *Meloidogyne* could not be resolved reliably, however, both NIP and sequence-based analyses suggest *Meloidogyne* to result from the more basal split (Figure 5) in concordance with another phylogenomic study (Philippe et al., 2004), but contradicting Lartillot and Philippe (2008) and Mortazavi et al. (2010). Further studies will show which topology is the best-supported hypothesis. Also the relations within the genus *Caenorhabditis* could not be resolved as expected. Both the position of *Brugia* and the alternative phylogeny of *Caenorhabditis* are not significantly supported (5a–b, Table 1). Here, fast intron loss appears to distort phylogenetic inference.

6. We obtain from four trochozoan species only very few novel introns that can be used to resolve the phyla Mollusca and Annelida (Figures 4 and 6). A second problem was that *Aplysia* is highly under-represented within the ortholog dataset (Supplementary Figure S7). At least, NIPs propose a common clade of trochozoan taxa but fail to resolve the split into annelids and molluscs. Moreover, the Platyhelminthes do not group as sister to the trochozoans to build the clade of Spiralia. A corresponding constrained search, however, did not require significantly longer trees (Table 1). Possibly the speed of intron evolution was particularly slow during the early radiation of molluscs and annelids, so that the origin of spiralian phyla cannot be resolved using NIP markers.

5. Conclusions

Overall, our study demonstrates that near intron pair (NIP) data could be used to derive a working hypothesis of the metazoan phylogeny using MP as tree reconstruction method. In particular, the analysis of NIP characters appears superior to an approach based on simple intron presence/absence data. Corresponding tree searches using Dollo parsimony and Wagner parsimony resulted in topologies that are clearly worse than the NIP-based predictions, respectively, see Supplementary Figures S8–S9. Thus, NIPs could be added to the already available set of rare genomic change (RGC) characters useful for tree reconstruction as all branchings of the strict consensus tree with more than 80% bootstrap support (Figure 3) are consistent with currently published phylogenies (Figure 4).

However, this first NIP-based phylogenetic analysis of a large, ancient taxon has uncovered also some methodical weaknesses. First, taxa near the supposed root of the tree and evolutionary periods of very low gain of introns, concerning here e.g. the lineages of *Mnemiopsis*, *Amphimedon*, *Hydra*, *Aplysia*, *Trichoplax*, *Strongylocentrotus*, *Nematostella*, and *Saccoglossus* (Figure 6) pose an objective challenge for NIP-based phylogenies. The unusual branching of *Ixodes* might be similarly caused by the much smaller fraction of novel introns in this genera compared to all other ecdysozoans analyzed. Second, models for the evolution of NIPs are not available. Under these circumstances, the necessary implementation of maximum parsimony exaggerates LBA effects. Probably, this causes the placement of *Mnemiopsis* as sister to the Pancrustacea, and the assemblage of several diploblastic species. Third, the positions of the *Caenorhabditis* and the mosquito species as well as that of *Daphnia* as sister to only some insect species might be due to the very unequal rates of intron gain and loss in evolution (Carmel et al., 2007; Krauss et al., 2008). This might be a hindrance for usage of NIP characters in a phylogenetic analysis covering many terminal taxa with very different evolutionary distances from each other but appear not do disturb studies concerning evolutionary splits of more comparable deepnesses (Krauss et al., 2008; Lehmann et al., 2010; Niehuis et al., 2012).

However, rapid developments in high-throughput sequencing are adding more genome sequences of good quality that better cover metazoan diversity. These data will likely also improve NIP-based phylogenies in the near future.

6. Supplementary Material

Supplementary Figures S1–S10, Supplementary Tables S1–S3 are available at Molecular Phylogenetics and Evolution online (http://www.journals.elsevier.com/molecular-phylogenetics-and-evolution/), Supplementary Materials 1–2 (partial alignments, NIP character matrices and trees) are available at http://www.bioinf.uni-leipzig.de/publications/supplements/11-003. NIP data matrix and MP trees of this study are also available at TreeBase, http://www.treebase.org, under study accession no. S13351.

7. Funding

This work was supported by the Deutsche Forschungsgemeinschaft (KR2065/2 to VK and STA850/6 to PFS). The Deutsche Forschungsgemeinschaft had no role in the design or interpretation of the study.

8. Acknowledgements

We gratefully acknowledge the availability of the sequencing data of the not yet published genomes of *Aplysia californica*, *Capitella teleta*, *Danio rerio*, *Helobdella robusta*, *Heterorhabditis bacteriophora*, *Ixodes scapularis*, *Lottia gigantea*, *Mnemiopsis leidyi*, *Rhodnius prolixus*, *Saccoglossus kowalevskii*, *Schistosoma japonicum*, and *Schmidtea mediterranea*. We would like to thank for the insightful comments of three anonymous reviewers on a previous version of this manuscript.

References

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

Belinky, F., Cohen, O., Huchon, D., Feb 2010. Large-scale parsimony analysis of metazoan indels in protein-coding genes. Mol Biol Evol 27 (2), 441–451.

Bininda-Emonds, O. R., 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics 6, 156.

Bourlat, S. J., Nielsen, C., Economou, A. D., Telford, M. J., Oct 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Mol Phylogenet Evol 49, 23–31.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T. L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Carmel, L., Wolf, Y. I., Rogozin, I. B., Koonin, E. V., 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res 17, 1034–1044.

Cho, S., Jin, S. W., Cohen, A., Ellis, R. E., Jul 2004. A phylogeny of Caenorhabditis reveals frequent loss of introns during nematode evolution. Genome Res 14 (7), 1207–1220.

Churakov, G., Sadasivuni, M. K., Rosenbloom, K. R., Huchon, D., Brosius, J., Schmitz, J., Jun 2010. Rodent evolution: back to the root. Mol Biol Evol 27, 1315–1326.

Coghlan, A., Wolfe, K. H., Aug 2004. Origins of recently gained introns in Caenorhabditis. Proc Natl Acad Sci U S A 101 (31), 11362–11367.

Conant, G. C., Stadler, P. F., May 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol 26, 1155–1161.

Csurös, M., Holey, J. A., Rogozin, I. B., Jul 2007. In search of lost introns. Bioinformatics 23 (13), i87–i96.

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., Feb 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965–968.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., Giribet, G., Apr 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452 (7188), 745–749.

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

Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32 (5), 1792–1797.

Edgecombe, G., Giribet, G., Dunn, C., Hejnol, A., Kristensen, R., Neves, R., Rouse, G., Worsaae, K., Srensen, M., 2011. Higher-level metazoan relationships: recent progress and remaining questions. Organisms Diversity & Evolution 11, 151–172.

Farris, J. S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.

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

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Vogel, J., Searle, S. M., Jan 2011. Ensembl 2011. Nucleic Acids Res 39 (Database issue), D800–D806.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., Lake, J. A., Mar 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267, 1641–1643.

Hejnol, A., Nov 2010. A twist in time–the evolution of spiral cleavage in the light of animal phylogeny. Integr Comp Biol 50 (5), 695–706.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., Dunn, C. W., Dec 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci 276 (1677), 4261–4270.

Irimia, M., Roy, S. W., Mar 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res 36 (5), 1703–1712.

Jow, H., Hudelot, C., Rattray, M., Higgs, P. G., Sep 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol 19, 1591–1601.

Krauss, V., Pecyna, M., Kurz, K., Sass, H., Jan 2005. Phylogenetic mapping of intron positions: a case study of translation initiation factor eIF2gamma. Mol Biol Evol 22, 74–84.

Krauss, V., Thümmler, C., Georgi, F., Lehmann, J., Stadler, P. F., Eisenhardt, C., May 2008. Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. Mol Biol Evol 25 (5), 821–830.

Kriegs, J. O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J., Schmitz, J., Apr 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol 4 (4), e91.

Kriegs, J. O., Zemann, A., Churakov, G., Matzke, A., Ohme, M., Zischler, H., Brosius, J., Kryger, U., Schmitz, J., Dec 2010. Retroposon insertions provide insights into deep lagomorph evolution. Mol Biol Evol 27 (12), 2678–2681.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G., Nov 2007. Clustal W and clustal X version 2.0. Bioinformatics 23 (21), 2947–2948.

Lartillot, N., Lepage, T., Blanquart, S., Sep 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Lartillot, N., Philippe, H., Apr 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci 363 (1496), 1463–1472.

Lehmann, J., Eisenhardt, C., Stadler, P. F., Krauss, V., 2010. Some novel intron positions in conserved Drosophila genes are caused by intron sliding or tandem duplication. BMC Evol Biol 10 (1), 156.

Long, M., Deutsch, M., Nov 1999. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol Biol Evol 16 (11), 1528–1534.

Mallatt, J., Craig, C. W., Yoder, M. J., Apr 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. Mol Phylogenet Evol 55 (1), 1–17.

Meusemann, K., von Reumont, B. M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J. W., Misof, B., Nov 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol 27 (1), 2451–2464.

Mortazavi, A., Schwarz, E. M., Williams, B., Schaeffer, L., Antoshechkin, I., Wold, B. J., Sternberg, P. W., Dec 2010. Scaffolding a Caenorhabditis nematode genome with RNA-seq. Genome Res 20, 1740–1747.

Nesnidal, M. P., Helmkampf, M., Bruchhaus, I., Hausdorf, B., Sep 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol 27, 2095–2104.

Nguyen, H. D., Yoshihama, M., Kenmochi, N., 2006. Phase distribution of spliceosomal introns: implications for intron origin. BMC Evol Biol 6, 69.

Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V., Eisenhardt, C., Hertel, J., Petersen, M., Mayer, C., Meusemann, K., Peters, R. S., Stadler, P. F., Beutel, R. G., Bornberg-Bauer, E., McKenna, D. D., Misof, B., Jul 2012. Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera. Current Biology 22 (14), 1309–1313.

Penny, D., McComish, B. J., Charleston, M. A., Hendy, M. D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53, 711–723.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T., Manuel, M., Wörheide, G., Baurain, D., Mar 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9, e1000602.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W., Casane, D., Sep 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21 (9), 1740–1752.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F., 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5, 50.

Pick, K. S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D. J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., Wörheide, G., Sep 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27, 1983–1987.

Prager, E. M., Wilson, A. C., 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J Mol Evol 27 (4), 326–335.

Qiu, W. G., Schisler, N., Stoltzfus, A., Jul 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol 21 (7), 1252–1263.

Rogozin, I. B., Lyons-Weiler, J., Koonin, E. V., Oct 2000. Intron sliding in conserved gene families. Trends Genet 16 (10), 430–432.

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., Koonin, E. V., Jun 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform 6, 118–134.

Roy, S. W., Gilbert, W., Mar 2005. Resolution of a deep animal divergence by the pattern of intron conservation. Proc Natl Acad Sci U S A 102 (12), 4403–4408.

Roy, S. W., Irimia, M., Apr 2008. Rare genomic characters do not support Coelomata: intron loss/gain. Mol Biol Evol 25 (4), 620–623.

Saeys, Y., Rouzé, P., Van de Peer, Y., Feb 2007. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. Bioinformatics 23 (4), 414–420.

Sankoff, D., 12 1974. Minimal mutation trees of sequences. SIAM Journal on Applied Mathematics 28 (1), 35–42.

Sankoff, D., Rousseau, P., 1975. Locating the vertices of a steiner tree in an arbitrary metric space. Mathematical Programming 9 (1), 240–246.

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

Savill, N. J., Hoyle, D. C., Higgs, P. G., Jan 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157, 399–411.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H. J., Hadrys, H., Dellaporta, S. L., Kolokotronis, S. O., Desalle, R., Jan 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol 7 (1), e20.

Slater, G. S., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., Plachetzki, D. C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S. F., Goodstein, D. M., Harris, C., Jackson, D. J., Leys, S. P., Shu, S., Woodcroft, B. J., Vervoort, M., Kosik, K. S., Manning, G., Degnan, B. M., Rokhsar, D. S., Aug 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466 (7307), 720–726.

Stamatakis, A., Nov 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22 (21), 2688–2690.

Swofford, D. L., 2003. PAUP* portable.

Talavera, G., Castresana, J., Aug 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56 (4), 564–577.

Templeton, A. R., 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37 (2), 221.

Trautwein, M. D., Wiegmann, B. M., Beutel, R., Kjer, K. M., Yeates, D. K., 2012. Advances in insect phylogeny at the dawn of the postgenomic era. Annu Rev Entomol 57, 449–468.

Weir, M., Eaton, M., Rice, M., 2006. Challenging the spliceosome machine. Genome Biol 7 (1), R3.

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

Wilkerson, M. D., Ru, Y., Brendel, V. P., Nov 2009. Common introns within orthologous genes: software and application to plants. Brief Bioinform 10 (6), 631–644.

Wolf, Y. I., Rogozin, I. B., Koonin, E. V., Jan 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14, 29–36.

Zdobnov, E. M., Bork, P., Jan 2007. Quantification of insect genome divergence. Trends Genet 23, 16–20.

Zheng, J., Rogozin, I. B., Koonin, E. V., Przytycka, T. M., Nov 2007. Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. Mol Biol Evol 24 (11), 2583–2592.

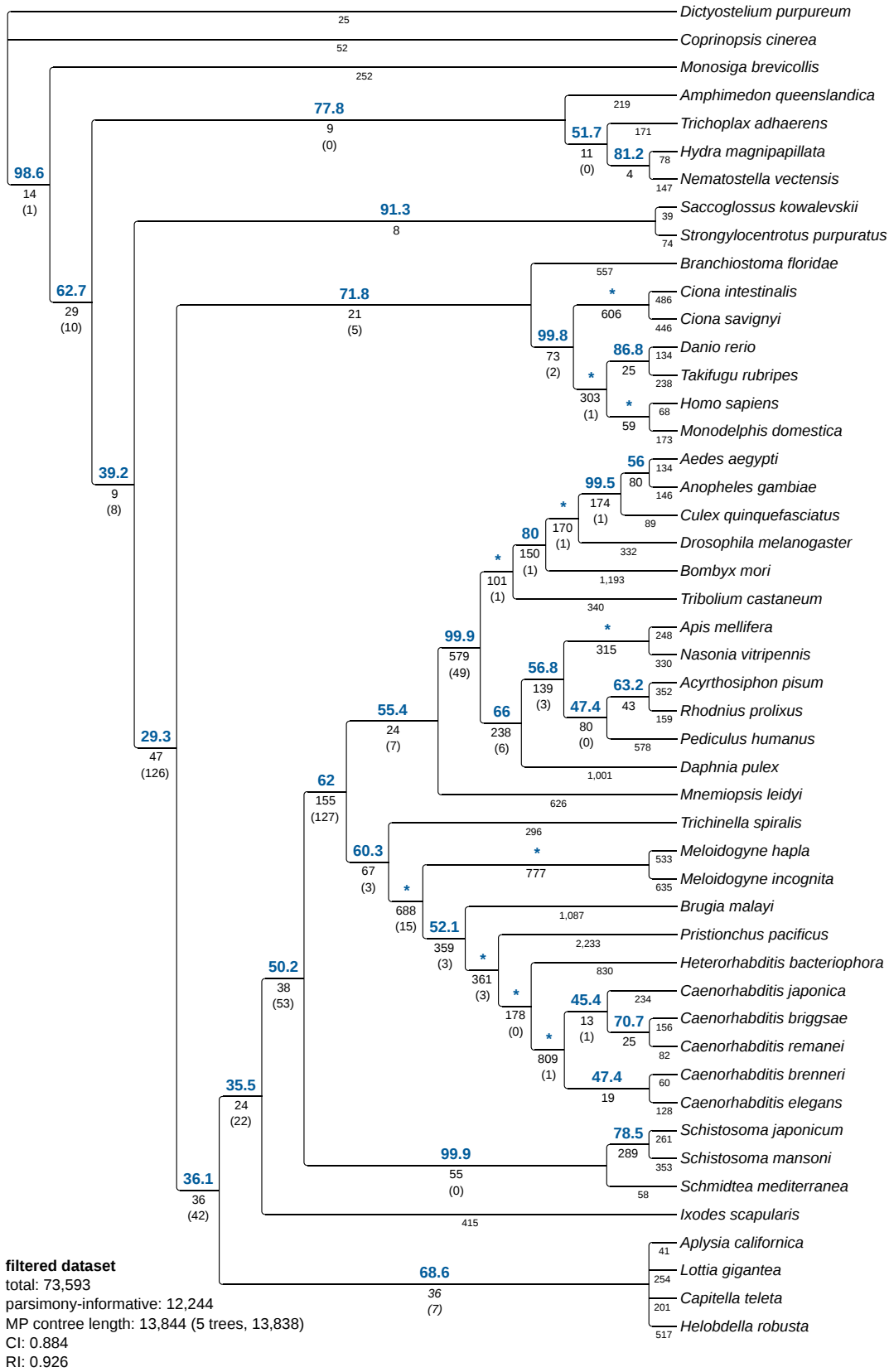

Figure 3: MP strict consensus tree. Bootstrap support (in percent) is given above the corresponding branches (in blue). A star denotes 100% support. In addition, the number of NIP characters that support the branch are shown below each branch. In parentheses, the number of characters inconsistent with the branch is given if possible. Note that in case of multifurcations, counts in italics may include cases with empty sub clades. All possible types of character distributions are illustrated in Supplementary Figure S1. The amoebozoan *Dictyostelium purpureum* was used to root the tree.

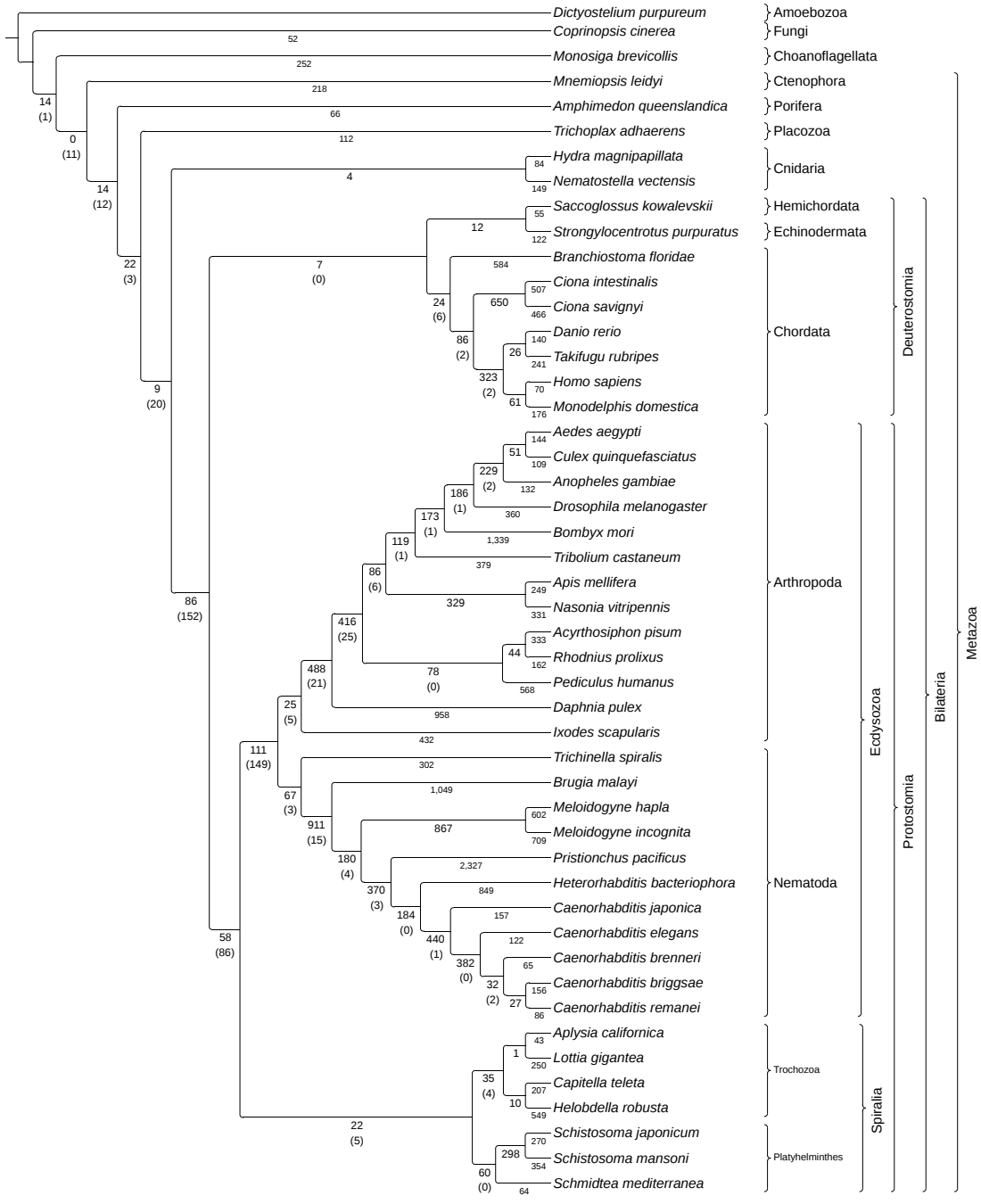

Figure 4: MP reconstruction on a majority consensus supertree of 45 analyzed metazoan species extracted from Lartillot and Philippe (2008), Hejnol et al. (2009), Mortazavi et al. (2010), and Meusemann et al. (2010). *Takifugu rubripes*, *Monodelphis domestica*, and *Heterorhabditis bacteriophora*, *Monosiga brevicollis*, *Coprinopsis cinerea*, and *Dictyostelium purpureum* were added according to their taxonomic position. The number of branch-supporting NIP characters is specified below each branch. Numbers in parentheses refer to characters that are inconsistent with the tree.

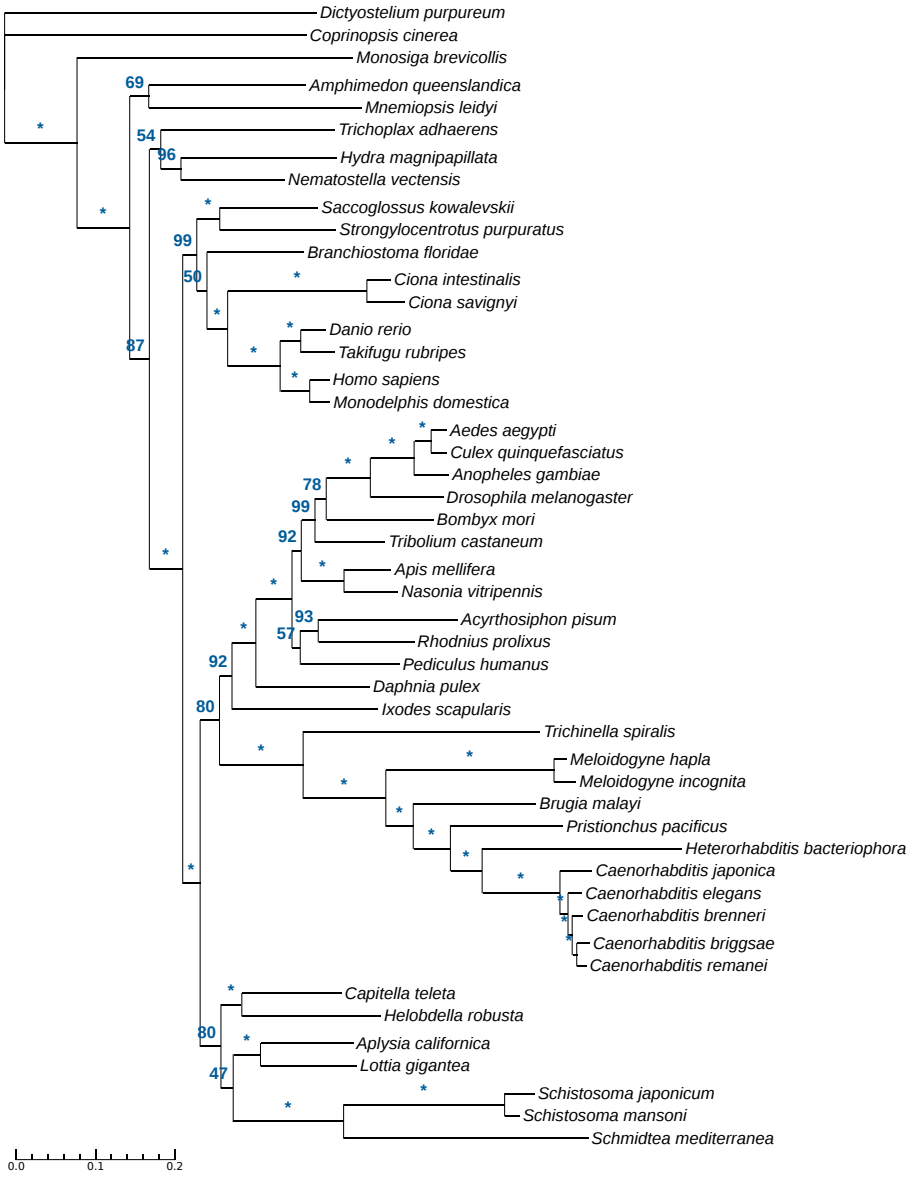

Figure 5: ML bootstrap consensus tree as obtained from conserved amino acid alignment columns and `RAxML`. Bootstrap percentages are given above the corresponding branches. Such with 100% support are indicated by a star.

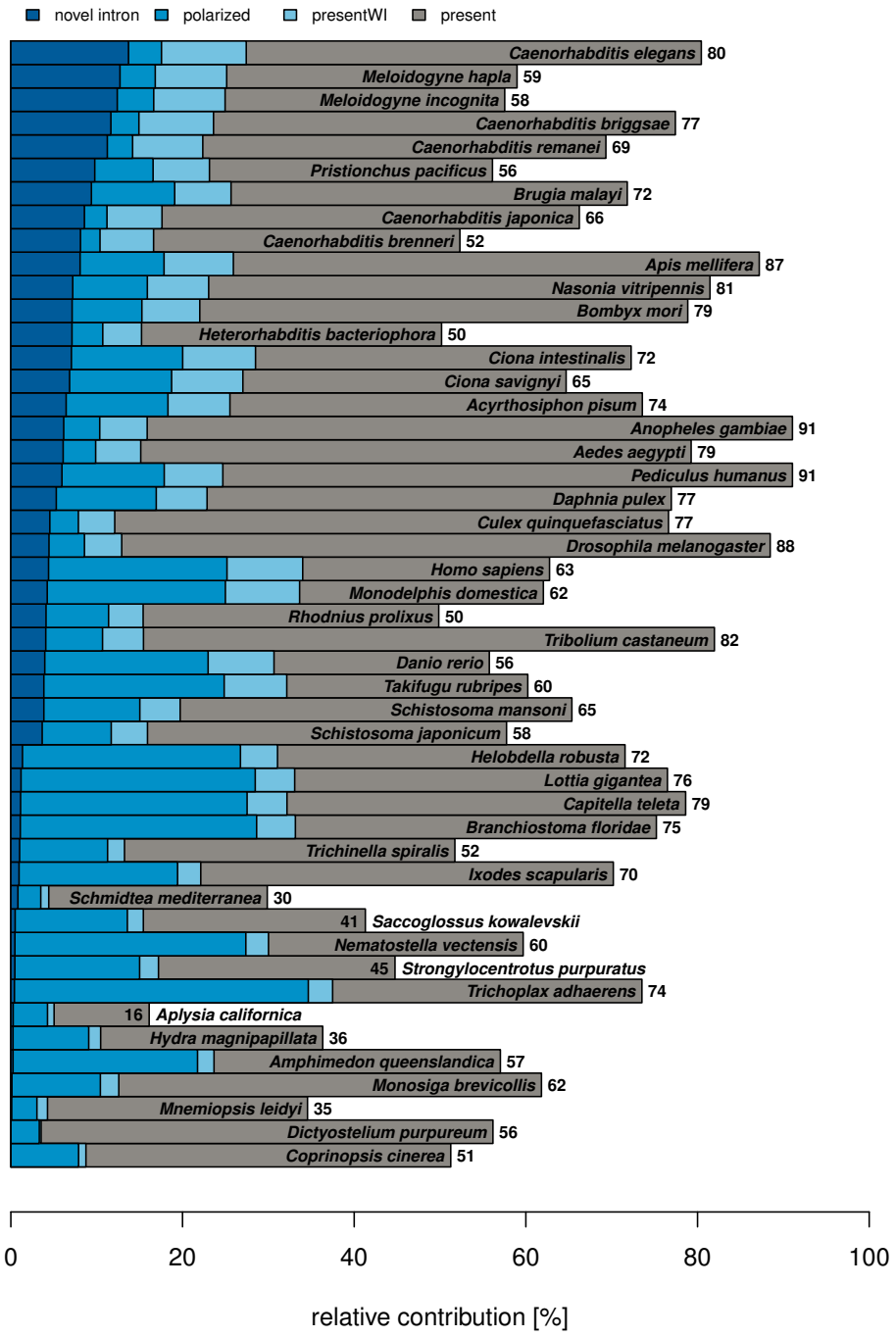

Figure 6: Contribution of individual taxa to the 12,244 parsimony-informative NIPs. The gray bars indicate the number of NIPs in which a taxon is present in the corresponding partial alignments ('present', percentages next to taxon names). The fraction in which a species additionally contributes an intron position to a NIP is indicated in light blue ('presentWI'). For the medium-blue colored subset of these NIPs, the characters could be polarized using the metazoan topology from Figure 4. Finally, taxa are sorted according to the percentage of NIPs to which they contribute the younger intron position (dark blue, 'novel intron').