Large-Scale Parsimony Analysis of Metazoan Indels in Protein-Coding Genes

Frida Belinky,[1] Ofir Cohen,[2] and Dorothée Huchon*[,1]

[1]Department of Zoology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv, Israel

[2]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv, Israel

***Corresponding author:** E-mail: huchond@post.tau.ac.il.

Associate editor: Hervé Philippe

Abstract

Insertions and deletions (indels) are considered to be rare evolutionary events, the analysis of which may resolve controversial phylogenetic relationships. Indeed, indel characters are often assumed to be less homoplastic than amino acid and nucleotide substitutions and, consequently, more reliable markers for phylogenetic reconstruction. In this study, we analyzed indels from over 1,000 metazoan orthologous genes. We studied the impact of different species sampling, ortholog data sets, lengths of included indels, and indel-coding methods on the resulting metazoan tree. Our results show that, similar to sequence substitutions, indels are homoplastic characters, and their analysis is sensitive to the long-branch attraction artifact. Furthermore, improving the taxon sampling and choosing a closely related outgroup greatly impact the phylogenetic inference. Our indel-based inferences support the Ecdysozoa hypothesis over the Coelomata hypothesis and suggest that sponges are a sister clade to other animals.

Key words: indel, phylogeny, indel-coding method, long-branch attraction, Ecdysozoa, Coelomata.

Research article

Introduction

Animals with a body cavity lined continuously by mesodermal derivatives (i.e., coelom), such as chordates and arthropods, were classically grouped in the Coelomata clade, whereas animals without a body cavity (Acoelomata) or with a body cavity only partially lined by mesodermal derivatives (Pseudocoelomata), such as nematodes, were considered to be separate lineages (Raff 1996; Halanych 2004). This view was challenged by 18S rRNA analyses that grouped molting animals (e.g., nematodes and arthropods) in the Ecdysozoa clade, altering the common view on animal evolution (Aguinaldo et al. 1997). Since the proposal of this alternative view, most molecular studies (Philippe et al. 2005, 2009; Irimia et al. 2007; Lartillot et al. 2007; Dunn et al. 2008; Roy and Irimia 2008; Schierwater et al. 2009) as well as morphological studies (Glenner et al. 2004) have shown support for Ecdysozoa. Conversely, others claim that the Coelomata topology is the correct one (Blair et al. 2002; Philip et al. 2005; Rogozin et al. 2007a, 2007b). Insertion and deletion (indel) analysis was also used to resolve the Ecdysozoa versus Coelomata debate. Wolf et al. (2004) analyzed indels from 384 orthologous alignments using Wagner parsimony and found a strong support in favor of Coelomata. Irimia et al. (2007), however, found three "clear cases of lineage-specific multiple amino acid indels" (Irimia et al. 2007, p 1606, fig. 4) in support of Ecdysozoa, although no such indels were presented in support of Coelomata.

Another controversy in animal phylogeny concerns the conflicting signals at the base of the metazoan tree. Although some sequence analyses place Porifera (sponges) as a sister clade to other metazoans (Borchiellini et al. 2001; Medina et al. 2001; Philippe et al. 2009), others, based on different data sets, group Porifera and Cnidaria (e.g., corals, medusas, sea anemones, and hydras) in the same clade (Dellaporta et al. 2006; Haen et al. 2007; Dunn et al. 2008; Schierwater et al. 2009). Clearly, resolving these deep animal relationships is a difficult task due to the limited number of sequences available for nonbilaterian metazoans (Baurain et al. 2007) and possibly also due to a rapid radiation early on in animal evolution (Rokas et al. 2005). To date, no indel analysis has been conducted in order to solve these deep metazoan relationships.

Indel analysis has been recently recognized as a powerful technique to resolve difficult phylogenies. The strength of the methodology stems from the assumption that homoplasy of indels is minimal, as independent insertions and deletions in the same position are considered unlikely (Rokas and Holland 2000; Bapteste and Philippe 2002). Recent interest in indel methodology has been triggered by the huge increase in indel data originating from genome projects. As a case in point, indels were successfully used to reconstruct bacterial phylogeny (Gupta 2001). A few variants of the indel methodology exist. For example, it is well known that short indels occur much more frequently than longer ones and hence the probability of homoplasy is much higher for short indels. It was consequently suggested that only multiresidue indels should be used for phylogenetic reconstruction (Lloyd and Calder 1991). In addition, there are several indel-coding approaches, two of which have been suggested as the best methods: simple indel coding (SIC) and the modified complex indel coding (MCIC)

(Simmons et al. 2007). In SIC (Simmons and Ochoterena 2000), each indel receives a separate two-state character of presence/absence. Any overlapping indels that exceed the boundaries of this indel are scored as missing data for that indel character. MCIC differs from SIC only in the treatment of overlapping indels (Müller 2006). MCIC uses multistate characters to code overlapping indels and assigns a distinct symmetrical step matrix to those gaps. More specifically, MCIC requires three steps. The first step is the delimitation of the characters. In the original definition of MCIC, each character was defined as a region of the alignment that was fully represented by one indel, and this indel was the longest existing one in this area (Müller 2006). The MCIC algorithm was recently improved so that each character is now a region of the alignment that contains all overlapping indels within it and not just the longest one (Müller K, personal communication). In the second step, the different states of each indel character are defined. Each sequence presenting a different indel pattern at the corresponding character region is coded as a different state. The third step is the determination of the number of steps between every two-character states. Each pair of sequences is compared separately for the corresponding character area, and the minimum number of steps between every two-character states is then determined. The MCIC approach thus aims at maximizing the information present in overlapping gaps because, unlike SIC, it does not require missing data to code overlapping gaps. In a recent study, it has been found that MCIC slightly outperforms SIC (Simmons et al. 2007).

In this study, we assessed the impact of these methodological variants on the above-mentioned debated metazoan relationships. Additionally, we studied the effects of taxon sampling on the inferred phylogeny. Finally, we compared the same analysis procedure on two independent ortholog data sets in order to assess the impact of ortholog prediction on our results. We thus performed a large-scale parsimony analysis of indels using various combinations of species, data sets, indel lengths, and coding methods.

Materials and Methods

Orthology Data sets

In order to address the impact of different orthology data sets on the inferred indel tree, we performed the same analyses on two independent ortholog data sets. One data set is based on the ComparaMart database of Ensembl (Flicek et al. 2008, ftp://ftp.ensembl.org/pub/release-49/mysql/compara_mart_homology_49/), the other is based on the eukaryotic orthologous groups of proteins (KOG) homology database (Tatusov et al. 2003, ftp://ftp.ncbi.nih.gov/pub/COG/KOG/). The Ensembl database was chosen because it has been described as one of the most reliable orthology databases (Flicek et al. 2008). The KOG database was chosen because it is the database used by Wolf et al. (2004) in their analyses showing that indels support the Coelomata hypothesis.

KOG-Based Data Set

Of the available species in the KOG database, those used to form an initial data set were *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. Only groups that contained exactly one homolog for each of the three animal species (i.e., *C. elegans*, *D. melanogaster*, and *H. sapiens*) were considered. This was done in order to avoid paralogs.

Ensembl-Based Data Set

Of the species available in the comparaMart database, the ones used to form the initial data set were *C. elegans*, *D. melanogaster*, *Anopheles gambiae*, *Ciona intestinalis*, *H. sapiens*, *Mus musculus*, and *S. cerevisiae*. We extracted all pairs of "one2one" orthologs of the above species (i.e., sequences from two taxa that are closer to each other than to any other sequence of the corresponding taxa). The one2one orthologs were then clustered into orthology groups. Specifically, if A was orthologous to B and B was orthologous to C, then A, B, and C were grouped together in the same cluster. To avoid paralogs, only groups that contained exactly one homolog for each of the six animal species (i.e., *C. elegans*, *D. melanogaster*, *A. gambiae*, *C. intestinalis*, *M. mus*, and *H. sapiens*) were considered.

Extending the KOG and Ensembl-Based Data Sets

The proteomes of 11 species were used to extend the KOG and Ensembl-based data sets. Protein sequences of *M. musculus*, *Danio rerio*, *Strongylocentrotus purpuratus*, *Brugia malayi*, *Nematostella vectensis*, and *Monosiga brevicollis* were retrieved from the National Center for Biotechnology Information (NCBI) protein database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein). Protein sequences of *A. gambiae* and *C. intestinalis* were retrieved from the Ensembl database (http://www.ensembl.org/info/data/ftp/index.html). DNA contigs of *Acropora millepora* were retrieved from Prof. Matz's web site (http://www.bio.utexas.edu/research/matz_lab/matzlab/454.html). Expressed sequence tag (EST) sequences of *Hydra magnipapillata* were downloaded from the NCBI EST database, and EST sequences of *Amphimedon queenslandica* were retrieved from the NCBI TRACE database (ftp://ftp.ncbi.nih.gov/pub/TraceDB/reniera_sp__jgi-2005). EST sequences of *H. magnipapillata* and *A. queenslandica* were assembled using the program CAP3 (Huang and Madan 1999). Protein-coding genes were then predicted from the contigs of *A. millepora*, *H. magnipapillata*, and *A. queenslandica* using the program ESTSCAN (Iseli et al. 1999).

Sequences from the 11 taxa mentioned above were added to the KOG and the Ensembl-based data sets using two consecutive BlastP searches. In the first search, for each additional species, the human representative of each orthology group was used as query against all available peptide sequences of this species. In the second search, all best matches obtained in the first search were used as query against the entire RefSeq database of human proteins (Pruitt et al. 2007). Only reciprocal best hits were considered to be orthologs. Our approach is a modification of the

classical reciprocal Blast search, in which the data set used as query in the first search is used as the target data set in the reciprocal search. Here, instead, the second search was performed against all available human protein sequences rather than only against the human sequences present in the KOG/Ensembl-based orthology data sets. The orthology groups used as query in the first Blast search represent only a fraction of the human proteome (1,029 groups for the KOG-based data set and 1,185 groups for Ensembl-based data set). Because for some of the added species only a part of the proteome has been sequenced, the best hit of the first search might be a paralogous sequences, that is, when the corresponding ortholog has not yet been sequenced. We found that using all known human protein sequences rather than just those of the original query for the reciprocal Blast, excluded such paralogs, because in the second search, the best hit would be another human protein absent from the original data set used for query. This approach is, in our case, more stringent and more reliable than a simple reciprocal Blast search.

Alignment Strategy and Alignment Quality Verification

Alignments were performed using amino acid sequences of the above-mentioned data sets. Each orthology group was aligned separately using ProbCons with default parameters (Do et al. 2005). The program Gblocks (Castresana 2000) was then applied to each group to remove poorly aligned positions, whereas retaining gaps within the remaining, well-aligned positions. The Gblocks parameters used were $-b1 = 75\%$ of the sequences present in the considered alignment (the minimum number of sequences for a conserved position), $-b4 = 5$ (the minimum length of a block), and $-b5 =$ a (enabling gaps in the output). Other parameters were set to default values. Each orthology group was aligned twice, once with the fungi sequences and once without. Because an erroneous alignment can lead to the comparison of nonhomologous indels, we assessed the quality of the alignments, after removing poorly aligned positions with Gblocks, by performing a reliability check using the "heads or tails" method (Landan and Graur 2007). To conduct the reliability check, sequence files after Gblock treatment were realigned twice (once without altering the order of the amino acid positions and once after reversing it). The results of these comparisons indicated that 78.2% of the orthology groups in the KOG-based data set and 86.5% of the orthology groups in the Ensembl-based data set had an alignment quality of over 95% (95% identical positions for all sequences in the alignment). Similarly, 99.4% and 100% of the orthology groups in the KOG and the Ensembl-based data sets, respectively, had a rate of identical residue pairs above 95% (the percentage of residue pairs that are paired identically in the two alignments). The distributions of quality scores, as well as identical residue pairs, are available as supplementary figure S1, Supplementary Materials online. Consequently, the vast majority of the orthology groups had high quality scores. Notably, these parameters are calculated based on conservation of amino acid positions and not indels. More specifically, even if the two alignments compared in the heads or tails method were different, the indel coding could still be identical. We verified that removing alignments with alignment quality below 95% did not change the topology obtained. For all the reasons indicated above, all alignments were retained in further analysis.

Sequence alignments of all orthology groups obtained after Gblocks treatment were concatenated. From these two concatenated alignments (with or without fungi sequences), five multiple sequence alignments (MSAs), differing in their represented species content, were derived. All MSAs included *H. sapiens*, *M. musculus*, *D. melanogaster*, *A. gambiae*, *C. elegans*, and *B. malayi*. Other species considered were MSA-1) fungi, *S. cerevisiae* (Ensemble-based data set) or *S. cerevisiae* and *S. pombe* (KOG-based data set); MSA-2) fungi, *M. brevicollis*, *N. vectensis*, *A. millepora*, *H. magnipapillata*, *A. queenslandica*, *S. purpuratus*, *C. intestinalis*, and *D. rerio*; MSA-3) *M. brevicollis*; MSA-4) *M. brevicollis*, *N. vectensis*, and *A. millepora*; MSA-5) *M. brevicollis*, *N. vectensis*, *A. millepora*, *H. magnipapillata*, *A. queenslandica*, *S. purpuratus*, *C. intestinalis*, and *D. rerio*. MSA-1 and MSA-2 were derived from the alignment including fungi; the other MSAs were derived from the alignment without fungi. These five MSAs are described in table 1.

Indel Coding and Tree Reconstruction

In classical sequence-based analysis, gaps are often treated as missing data; this is, for example, the default option in PAUP* (Swofford 2003). Similarly, alignment programs do not differentiate between gaps and missing data; however, in indel analysis, it is imperative to distinguish between indels and missing data. Consequently, gaps starting at the N-terminus or ending at the C-terminus of the alignment were coded as missing data. Additionally, it is important to exclude gaps that might be the result of erroneous prediction of splice sites. We took a conservative approach as we arbitrarily a priori excluded all gaps longer than 50 amino acids from the analysis because such long deletions in conserved protein-coding genes are most probably the results of artifacts. Additionally, it has been shown that the average exon length, in model eukaryotes, is ~25 amino acid long and that the vast majority of exons are longer than five amino acids (Deutsch and Long 1999; Yandell et al. 2006). Consequently, gaps shorter than five amino acids are not likely to be the result of erroneous splice site predictions. We thus verified that 97% of the informative indels contained in our data set were shorter than or equal to five amino acids.

Indels contained in the concatenated data sets were coded using either the MCIC or the SIC method as implemented in the program SeqState v1.4.1 (Müller 2005). For each of the 10 MSAs (five KOG-based and five Ensembl-based MSAs), and for each of the two coding methods, two possibilities were considered: 1) including all indels and 2) excluding single-residue indels. Exclusion of single-residue indels was performed by replacing single-residue

Table 1. Comparison between Coelomata and Ecdysozoa Topologies.

Data Set		Ensembl-Based Data Set				KOG-Based Data Set			
Indel Length		All Indels		Multiresidue Indels Only		All Indels		Multiresidue Indels Only	
Indel-Coding Method		SIC	MCIC	SIC	MCIC	SIC	MCIC	SIC	MCIC
MSA-1:	Best	C	C	C	C	C	C	C	C
1–2 Fungi*	BP	100	97	100	100	100	95	100	100
6 Bilateria	Test	C	C	C	C	C	C	C	C
	P value	<0.0001	0.001	<0.0001	0.0001	<0.0001	<0.0001	<0.0001	<0.0001
MSA-2:									
1–2 Fungi*	Best	O	O	O	O	O	E	O	O
1 Choanozoa	BP	41	49	50	46	49	64	11	17
1 Porifera	Test	—	—	—	—	—	—	—	—
3 Cnidaria	*P* value	0.026	0.011	0.396	0.037	0.028	0.020	0.796	0.564
9 Bilateria									
MSA-3:	Best	C	C	C	C	C	C	C	C
1 Choanozoa	BP	61	59	64	75	55	49	84	83
6 Bilateria	Test	—	—	—	—	—	—	—	—
	P value	0.742	0.752	0.732	0.493	0.706	0.895	0.275	0.346
MSA-4:	Best	E	E	E	E	E	E	O	O
1 Choanozoa	BP	85	82	77	60	74	82	63	64
2 Cnidaria	Test	E	E	—	—	E	E	E	E
6 Bilateria	*P* value	<0.0001	0.0001	0.016	0.034	<0.0001	<0.0001	0.0006	0.0006
MSA-5:	Best	E	E	E	E	E	E	E	E
1 Choanozoa	BP	94	84	79	68	82	85	42	28
1 Porifera	Test	E	E	E	E	E	E	—	—
3 Cnidaria	*P* value	<0.0001	<0.0001	0.0001	0.0001	0.0007	0.001	0.050	0.018
9 Bilateria									

C—Coelomata, E— Ecdysozoa, O—other (usually, nematodes sister clade to all other animals); Best—MP topology; BP—bootstrap percentage of the corresponding supported clade (C, E, or O); Test—winning topology between Coelomata and Ecdysozoa topologies based on Templeton test; *P* value—significance of the Templeton test; significant *P* values after applying a Bonferroni correction are underlined. MSA—multiple sequence alignment. *The Ensembl-based data sets include one fungus sequence, whereas the KOG-based data sets include two fungi sequences.

gaps with an "X". Thus, single-residue indels are converted to missing amino acid characters and hence not treated as indels. To compare the homoplasy present within multiresidue with the homoplasy present within single-residue indels, data sets excluding multiresidue indels were created by replacing all multiresidue gaps with Xs. It is worth noting that due to our masking procedure and the way in which overlapping indels are coded, the total number of all indels differs from the sum of multiresidue and single-residue indels. Because overlapping indels can include both single and multiresidue indels, MCIC, for example, will code such indels as one character in all three data sets (i.e., all indels, multiresidue indels only, and single-residue indels only). The number of indels and informative indels present in each of the 50 data sets is indicated in supplementary tables S1–S4, Supplementary Materials online. The data sets are available on our website http://www.tau.ac.il/~huchond/Supplementary/Belinky_metazoan_indel.html.

Maximum parsimony (MP) reconstructions were performed using PAUP* v4.0b10 (Swofford 2003). Best tree searches were performed with the branch-and-bound algorithm. Bootstrap analyses (Felsenstein 1985) were conducted using heuristic searches with the tree bisection and reconnection (TBR) branch–swapping option and with 100 random addition sequences. Notably, we verified that best tree searches performed with TBR branch swapping gave the same results as branch-and-bound searches. Bootstrap percentages (BPs) were computed after 500 replicates. Bremer indices provided similar information as the bootstrap supports and were thus not presented.

Hypothesis Testing

For each of the 40 multiresidue and all indel data sets, the Templeton test (Templeton 1983), using two-tailed probabilities as implemented in PAUP*, was used to evaluate whether the MP tree under the Coelomata hypothesis significantly differs from the MP tree under the Ecdysozoa hypothesis. In each analysis, these two topologies were compared regardless of the best tree obtained in the MP search described above. The Templeton test was also used to determine whether a first divergence of Porifera among Metazoa is significantly more parsimonious than a sister–clade relationship of Porifera and Cnidaria. A Bonferroni correction was applied to account for multiple testing.

Results and Discussion

Our design allowed us to test the impact of four factors on indel-based phylogeny: the taxonomic sampling, the data set, the inclusion/exclusion of single-residue indels, and the indel-coding method.

Impact of Taxonomic Sampling

Among all parameters tested, taxonomic sampling was found to have the greatest impact on indel-based phylogenetic inferences. The impact of taxonomic sampling

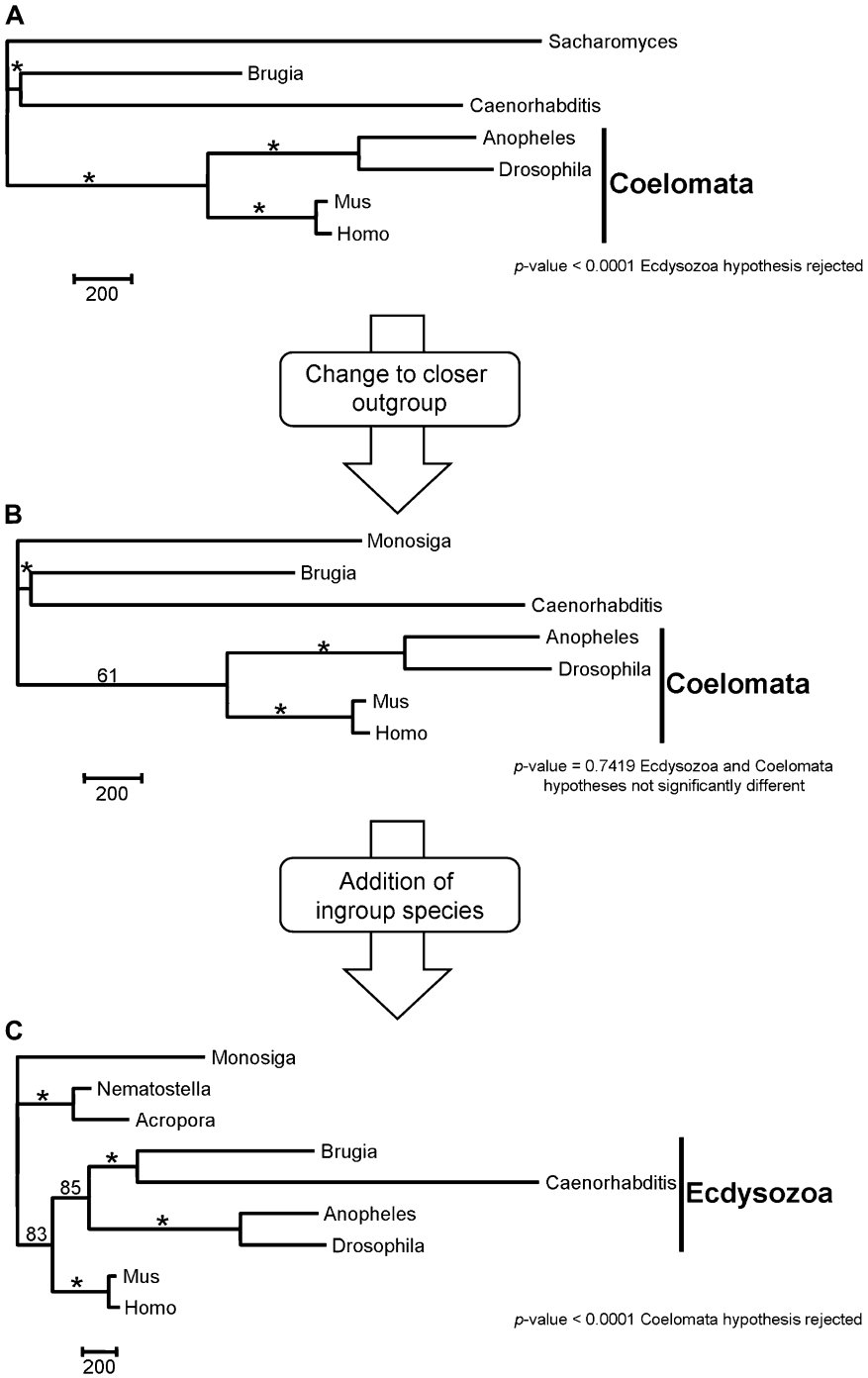

Fig. 1. MP phylogenetic trees, reconstructed based on the SIC method, including all indel lengths and using the Ensembl-based data set. (*A*) outgroup: *Saccharomyces cerevisiae*. (*B*) outgroup: *Monosiga brevicollis*. (*C*) outgroup: *M. brevicollis*, *Nematostella vectensis*, and *Acropora millepora*. BPs are given above the corresponding branches. Branches with BP = 100 are indicated with a star. *P* values of Templeton test comparing between the Coelomata and the Ecdysozoa hypothesis are given below each tree.

on metazoan phylogeny was tested using different combinations of outgroup and nonbilaterian metazoan species. All indel analyses that comprised only six animal taxa-two mammals, two insects, and two nematodes, with fungi as outgroup (i.e., MSA-1), resulted in a highly supported Coelomata topology regardless of the data set, indel length, or indel-coding method used (BP = 94–100; fig. 1*A* and supplementary fig. S2, Supplementary Material online). Statistical comparisons between the Coelomata and Ecdysozoa topologies for MSA-1 significantly support Coelomata (table 1, MSA-1, Templeton test *P* value $\leq$ 0.001). This result is in agreement with the analysis of Wolf et al. (2004).

If indels indeed support Coelomata, the same result is expected when rooting with a closer outgroup. Choanoflagellates have been found to be the closest relative of animals (King et al. 2008). However, changing the outgroup to the choanoflagellate *M. brevicollis* (i.e., MSA-3) greatly reduced the support for Coelomata to BP = 40–84 (fig. 1*B*

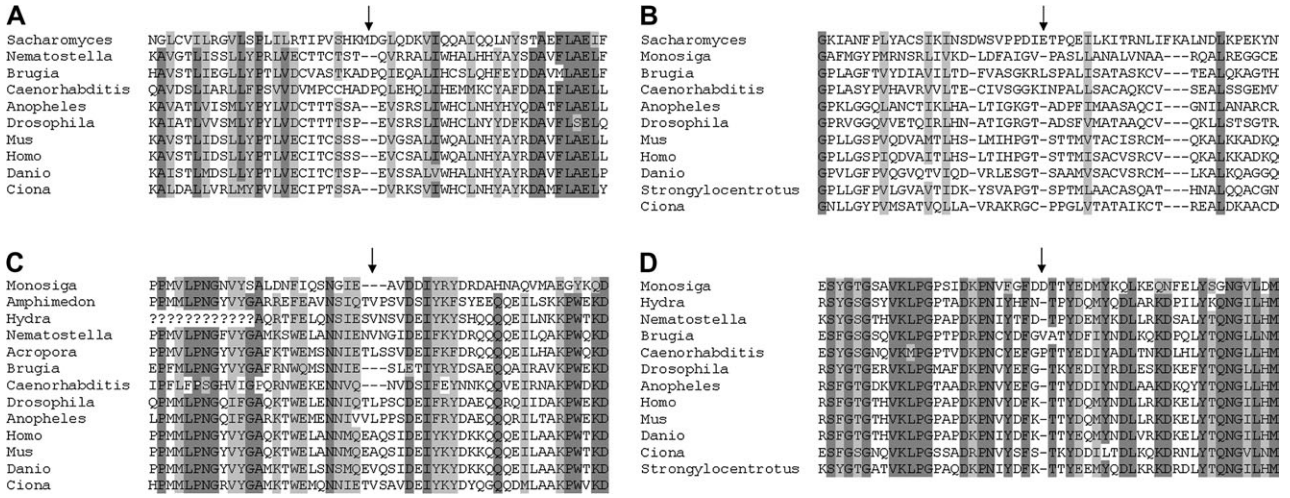

FIG. 2. Examples of homoplastic indels, present in the Ensembl-based data set, supporting Coelomata with a limited taxon sampling. (*A*,*B*) outgroup: *Saccharomyces cerevisiae*. (*C*,*D*) outgroup: *Monosiga brevicollis*. (*A*,*C*) multiresidue indels. (*B*,*D*) single-residue indels. The arrow indicates the location of the homoplastic indel.

and supplementary fig. S3, Supplementary Material online). Furthermore, the Coelomata and Ecdysozoa topologies are not statistically different in this case (table 1, MSA-3, Templeton test *P* value = 0.275–0.895). The addition of two more species (i.e., the cnidarians *N. vectensis* and *A. millepora*) shifts the topology in favor of Ecdysozoa in most analyses (table 1, MSA-4) with moderate support (BP = 61–85, fig. 1C and supplementary fig. S4, Supplementary Material online). Although the support values are moderate, in most cases, the Ecdysozoa topology is significantly better than the Coelomata one, based on Templeton tests (table 1, MSA-4). It is worth noting that the Templeton test is a conservative test when using two-tailed probabilities (Larson 1994). The addition of more metazoan taxa increases the support for Ecdysozoa up to BP = 94 (table 1, MSA-5, and supplementary fig., S5*E* Supplementary Material online). Thus, our results suggest that the Coelomata outcome is the result of a long-branch attraction (LBA) artifact (Felsenstein 1978), because Coelomata is only supported using the most distant outgroup and only when the species sampling is limited to six bilaterians. Notably, this result (i.e., a shift from Coelomata to Ecdysozoa after increasing the taxon sampling) is in agreement with conclusions from standard multigene analysis (Delsuc et al. 2005). To avoid LBA, in our analyses, the addition of taxa is obligatory yet insufficient, as analyses that use fungi as outgroup and include additional metazoan taxa support neither Coelomata nor Ecdysozoa (table 1, MSA-2, and supplementary fig. S6, Supplementary Material online). Only the combination of close outgroups and large ingroup sampling significantly shifts the topology toward Ecdysozoa.

The fact that species sampling has an impact on indel-based phylogeny indicates the presence of homoplasy in our data set. Homoplasy (i.e, the errors resulting from non-phylogenetic signals) is a well-known problem of sequence-based phylogeny (Baurain et al. 2007). Although homoplasy of indels has been considered before (Bapteste and Philippe 2002; de Jong et al. 2003), it has not yet been reported in large-scale indel analyses of protein-coding sequences. Indels have often been considered to be ideal markers for phylogeny (Lloyd and Calder 1991; Rokas and Holland 2000; Gupta and Mok 2007). Because of their scarcity in protein sequences when compared with substitutions, it has been assumed that indel homoplasy would be minimal. Our own results, on the other hand, show that homoplasy present within indel characters can lead to LBA. As is the case for substitution-based analyses (Lecointre et al. 1993; Graybeal 1998), species sampling has a strong impact on indel-based phylogeny. Examples of homoplastic indels are shown in figure 2. In figure 2*A* and *B*, identical indels are shared between fungi and nematodes but not by any other animal or choanoflagellate sequences. Consequently, when only fungi are considered together with bilaterian sequences (MSA-1), these indels support the Coelomata hypothesis. A larger species sampling (e.g., MSA-2) reveals that these indels are most likely homoplastic, in particular because they are absent in the choanoflagellate and/or cnidarians. However, it is important to note that even if the addition of taxa allows detection of homoplastic indels, it does not remove all conflicts in the data set. Indeed, indels in support of Coelomata are still present in MSA-5, although less frequent than Ecdysozoan indels, as indicated by the bootstrap values and Templeton test results (table 1). Examples of conflicting indels in support of Ecdysozoa and Coelomata are presented in supplementary figure. S7, Supplementary Materials online.

Impact of Data Set

Two data sets of putative orthologs were analyzed —a KOG-based data set and an Ensembl-based data set. The two data sets were used as reciprocal controls for potential differences in ortholog prediction. No differences between the two data sets were anticipated. However, although both data sets exhibit the same behavior in respect to changes in root selection, species sampling, and coding method, small differences between the phylogeny inferred

Table 2. CI and RI Values Calculated Based on Multiresidue and Single-Residue Indels.

Database		Ensembl-Based Data Set			KOG-Based Data Set		
Indel Length		Multiresidue Indels Only		Single-Residue Indels Only	Multiresidue Indels Only		Single-Residue Indels Only
Indel-Coding Method		SIC	MCIC	SIC	SIC	MCIC	SIC
	CI	0.97	0.98	0.94	0.97	0.97	0.94
MSA-1	RI	0.86	0.88	0.78	0.85	0.87	0.80
	CI	0.93	0.93	0.84	0.92	0.93	0.84
MSA-2	RI	0.68	0.58	0.58	0.65	0.66	0.60
	CI	0.96	0.97	0.93	0.96	0.97	0.92
MSA-3	RI	0.86	0.88	0.78	0.81	0.82	0.76
	CI	0.95	0.96	0.90	0.95	0.96	0.88
MSA-4	RI	0.81	0.82	0.72	0.76	0.70	0.70
	CI	0.92	0.93	0.84	0.91	0.92	0.82
MSA-5	RI	0.70	0.71	0.61	0.63	0.64	0.60

from each data set do exist (table 1). In most cases, the overall bootstrap support is higher in the Ensembl-based data set, in particular with MSA-4 and 5 (where the outgroup is the choanoflagellate *M. brevicollis*). Furthermore, the Ensembl-based data set seems to be more robust than the KOG-based data set to changes in indel length, because the same topology is recovered with similar bootstrap support independent of the inclusion or exclusion of single-residue indels. In particular, excluding single-residue indels provides a stronger support for the Coelomata hypothesis when considering the KOG-based data set, but not with the Ensembl-based data set. For example, different topologies are obtained using the MSA-4 of the KOG-based data set with and without single-residue indels. When only multiresidue indels are considered, the best topology is neither Coelomata nor Ecdysozoa, whereas when all indels are considered, the best topology becomes Ecdysozoa. Such topological changes do not exist with the Ensembl-based data set. When using the MSA-4 of the Ensembl-based data set, all indel and multiresidue analyses agree on the same topology (supplementary fig. S4, Supplementary Materials online). The lower robustness of the KOG-based data set can be explained by the smaller number of characters present in the KOG-based data set compared with the Ensembl-based data set (supplementary tables S1–S4, Supplementary Material online) or by the slightly lower quality of the alignments compared with the Ensembl-based data set. The difference in alignment quality between the data sets might result from a less accurate ortholog prediction (supplementary fig. S1, Supplementary Material online). Indeed, the KOG database has been suspected of having a higher level of erroneous predictions of orthologs compared with several other orthology databases including Inparanoid, which was evaluated as one of the best (Chen et al. 2007). Although the performance of the KOG database was not directly compared with Ensembl, it was recently reported that the orthology prediction of the Ensembl database outperforms Inparanoid (Vilella et al. 2009).

Impact of Inclusion/Exclusion of Single-Residue Indels

It has been suggested that multiresidue indels are more reliable than single-residue indels because they might be less prone to homoplasy (Lloyd and Calder 1991). We thus compared data sets with all indel lengths versus data sets with only multiresidue indels, in order to test the hypothesis that excluding single-residue indels would improve the inferred phylogeny. In particular, an increase in branch support was expected when excluding single-residue indels. Surprisingly, the opposite result was usually observed, as most trees obtained with all indels showed overall higher support than trees obtained only with multiresidue indels (table 1 and supplementary fig. S2–S6, Supplementary Material online). The lower bootstrap support obtained only with multiresidue indels might be due to a reduction in phylogenetic signals as single-residue indels are the most abundant in the data sets (supplementary tables S1–S4, Supplementary Material online). Phylogenetic results inferred from MSA-3 present an exception to this rule. In this case, the data support the Coelomata hypothesis and the support increases when single-residue indels are excluded (table 1). Because the Coelomata hypothesis is not supported when more species are added (MSA-4, 5), we cannot attribute this increase in support to a better phylogenetic signal contained in multiresidue indels. On the contrary, this suggests that not only single-residue indels but also multiresidue indels contain some level of homoplasy. Examples of probable homoplastic indels present in MSA-3 are presented in figure 2C and *D*. Consequently, our results support the idea that single-residue indels should not be a priori excluded from large-scale indel-based analysis (Simmons et al. 2001).

Homoplasy of Indel Characters

As described above, we found substantial homoplasy within indel characters. Examples of probable homoplastic indels present can be found in figure 2. Comparing the consistency index (CI) and retention index (RI) of multi versus single-residue indels reveals that the CI and RI of single-residue indels are always lower than that of multiresidue indels for the same data set (table 2). This reflects the greater amount of homoplasy within single-residue indels. This high level of homoplasy is also reflected in the inferred phylogeny based on single-residue indels only (supplementary fig. S8, Supplementary Material online), because in these trees the long-branched nematodes are often

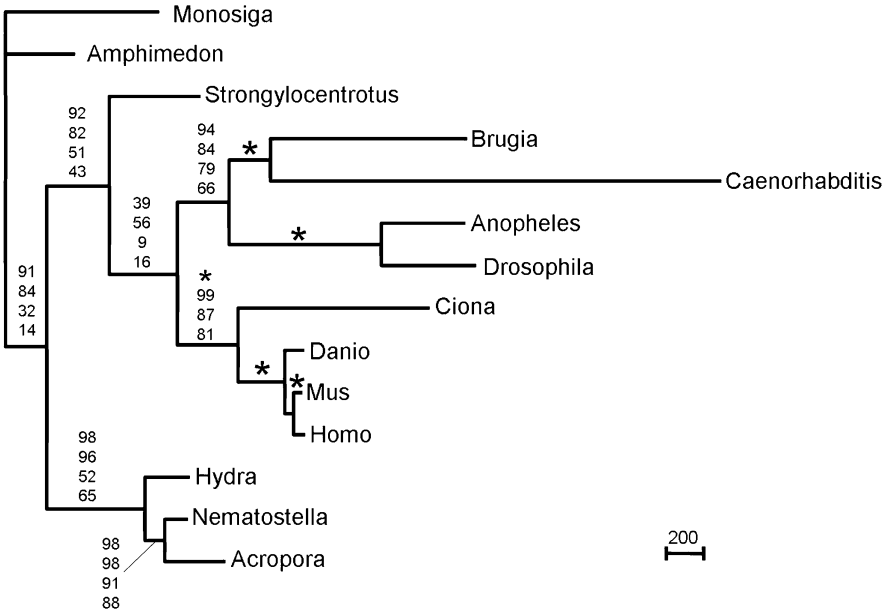

Fig. 3. MP tree (9,641 steps) based on all indel data coded using SIC from 1,185 predicted ortholog groups present in the Ensemble-based data set. Nodes with maximal BPs are indicated with a star. The four support values at each node represent, from top to bottom: (1) BP for all indels using SIC; (2) BP for all indels using MCIC; (3) BP for multiresidue indels using SIC; and (4) BP for multiresidue indels using MCIC.

attracted by the outgroup. As a case in point, in MSA-4, single-residue indel analysis suggests that nematodes diverged before cnidarians, whereas analyses of all indels or of multiresidue indels support the Ecdysozoa hypothesis (supplementary fig. S4 and S8, Supplementary Material online). However, single-residue indels do contain a nonnegligible phylogenetic signal that is revealed when species sampling is increased. It is remarkable that with the largest species sampling and without the long-branched fungi (i.e., MSA-5), the tree reconstructed based on single-residue indels only (supplementary fig. S8, Supplementary Material online) agrees with the tree reconstructed based on multiresidue indels (supplementary fig. S5, Supplementary Material online). This further supports the idea that single-residue indels should not be removed a priori from indel analyses (Simmons et al. 2001).

Impact of Coding Method

It has previously been suggested that SIC and MCIC are the best methods for indel coding, with MCIC slightly outperforming SIC (Simmons et al. 2007). In our results too, SIC and MCIC performed similarly: Among the 40 trees reconstructed, the phylogenetic trees reconstructed by the two coding methods only differed in lowly supported nodes (BP $<$ 75%). Generally, the bootstrap supports tend to be slightly higher for SIC, probably due to the fact that SIC provides matrices with more characters than MCIC due to different treatment of overlapping indels (supplementary tables S1–S4, Supplementary Material online).

Phylogenetic Inference

Coelomata versus Ecdysozoa. Depending on the data set considered, support was found in favor of either Coelomata or Ecdysozoa. The different conditions under which each of these hypotheses prevails are informative. In accordance with phylogenetic studies based on sequence substitutions, the Coelomata topology is the best topology under conditions which favor LBA: limited taxon sampling, distant outgroup, and heterogenous evolutionary rate among ingroup species (Philippe and Laurent 1998). Altering these conditions, by choosing a larger taxon sampling and a closer outgroup (MSA-4, 5), leads to phylogenetic trees that instead significantly support the Ecdysozoa phylogeny. This suggests that the Coelomata hypothesis is the result of LBA. Indeed, had the Coelomata topology been the true one, the use of a close outgroup and large taxon sampling should not have drastically affected the topology and its statistical support. When neither the Coelomata nor the Ecdysozoa topologies are recovered (mainly MSA-2, supplementary fig. S6, Supplementary Material online), the topology obtained places the nematodes as a sister clade of other metazoans, with low support (BP = 41–50). This result further supports the notion that in the presence of a distant outgroup (fungi are included in MSA-2), the nematodes are attracted to the base of the tree, which is consistent with the idea that the significant support observed for the Coelomata hypothesis in MSA-1 analysis is the result of LBA.

We thus conclude that, in contrast to Wolf et al. (2004), our indel-based analysis supports Ecdysozoa rather than Coelomata, in agreement with recent large-scale phylogenetic sequence analyses (Dunn et al. 2008; Philippe et al. 2009; Schierwater et al. 2009).

Animal Phylogeny. Based on the above results, animal relationships will only be discussed here for the MSA-5 analysis using the Ensembl-based data set, because this has been found to be the most reliable setting. The indel-based phylogenetic tree obtained agrees with sequence-based inferences (Philippe et al. 2009) (fig. 3 and supplementary fig. S5 *E–H*, Supplementary Material online).

Analyses based on all indels support a sister position of sponges to other animals (BP = 84–91). However, Templeton test, comparing a sister-group relationship of the sponges and cnidarians versus a sister position of sponges to other animals, reveals that the two topologies are not significantly different (MCIC *P* value = 0.166, SIC *P* value = 0.126). Moreover, multiresidue indels do not support a sister position of sponges to other animals. The reason for these inconclusive results is probably the large amount of missing data that exist in our data set for all nonbilaterian species (38–81% of missing data, supplementary tables S5–S6, Supplementary Material online). Indeed, very few indels were inferred to be unambiguous synapomorphies of each hypothesis: ten single-residue indels and two multiresidue indels were inferred in favor of a sister position of sponges to other animals, whereas five single-residue indels and two multiresidue indels were inferred in favor of a sister–clade relationship between sponges and cnidarians. Examples of such indels are presented in supplementary figure. S9, Supplementary Material online, which reveals, that these few multiresidue indels are not reliable characters. They are either located in regions of the alignment where the sponge sequence is poorly aligned or in regions that appear to be hotspots of indels. The indel-based inference of deep metazoan relationships is expected to become more reliable with the completion of the sponge genome project (Hooper and Van Soest 2006) and when more complete EST information from representatives of other sponge classes becomes available (i.e., Calcarea, Homoscleromorpha, and Hexactinellida).

Both cnidarians and bilaterians are found to be monophyletic. Among bilaterians, the monophyly of deuterostomes is not supported. In most cases, *S. purpuratus*, the sea urchin, is placed as sister clade to all other Bilateria although without support (BP < 40). It is worth noting that the monophyly of Deuterostomia is also not always recovered with primary sequences analysis (Lartillot and Philippe 2008); thus, the lack of a phylogenetic signal for Deuterostome monophyly is a common feature of both primary sequences and indel analyses. In contrast, indel analyses support the monophyly of chordates (BP = 85–100), whereas chordate monophyly is usually not recovered using mitochondrial sequences because the high evolutionary rate of tunicate mitochondrial genomes places them at a sister position to other Bilaterians (Bourlat et al. 2008). This indicates that although indels contain some level of homoplasy, large-scale indel analysis can be complementary to sequence-based analysis to help resolve debated phylogenetic relationships.

Conclusions

Although indels are often considered to be ideal markers with almost no homoplasy, our results indicate that their analysis should be treated carefully. Both single-residue and multiresidue indels appeared to contain a nonnegligible level of homoplasy and to be prone to LBA. Similar to sequence data, indel-based inference appears to be mainly sensitive to taxon sampling, whereas indel-coding methods, orthology data sets, or indel lengths appear to influence the phylogenetic inference to a lesser extent. Consequently, our results suggest that one should be careful when interpreting phylogenetic inference based on a few indel loci from a few genes (van Dijk et al. 1999; de Jong et al. 2003; Gupta 2006) or based on limited taxon sampling (Wolf et al. 2004). Because indel characters are rare in protein sequences compared with amino acid substitutions, large-scale parsimony analyses of indels are only applicable to genome sequences with sufficiently high coverage. This, in turn, limits the taxon sampling in indel analysis compared with classical sequence analysis (Philippe et al. 2009). In addition, the parsimony criterion used to infer the indel-based tree is known to be highly sensitive to LBA (Felsenstein 1978), which is a limitation compared with the advanced probabilitistic models used to analyze DNA and protein sequences (Lartillot and Brinkmann 2007). However, in spite of these limitations, we believe that indel information is an important complement to sequence-based phylogeny. It is expected that with time, more genomes with high coverage will become available. Furthermore, the implementation of likelihood models adapted to indels should improve indel-based phylogeny, even in the presence of distant outgroups.

Supplementary Material

Supplementary tables S1–S6 and supplementary figures S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Hervé Philippe, Tal Pupko, and three reviewers for their comments and helpful suggestions, Kai Müller for his help with the SeqState program, Naomi Paz for revising the English text, and Noam Ariel for assistance in indel data analysis. O.C. is a fellow of the Edmond J. Safra program in bioinformatics. This research was supported by the Israel Science Foundation (grant No. 600/06 to D.H).

References

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493.

Bapteste E, Philippe H. 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. *Mol Biol Evol*. 19:972–977.

Baurain D, Brinkmann H, Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Mol Biol Evol*. 24:6–9.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. *BMC Evol Biol*. 2:7.

Borchiellini C, Manuel M, Alivon E, Boury-Esnault N, Vacelet J, Le Parco Y. 2001. Sponge paraphyly and the origin of Metazoa. *J Evol Biol*. 14:171–179.

Bourlat SJ, Nielsen C, Economou AD, Telford MJ. 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. *Mol Phylogenet Evol*. 49:23–31.

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

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

de Jong WW, van Dijk MA, Poux C, Kappe G, van Rheede T, Madsen O. 2003. Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol Phylogenet Evol*. 28:328–340.

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci USA*. 103:8751–8756.

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

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*. 27:3219–3228.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*. 15:330–340.

Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

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

Felsenstein J. 1985. Confidence-limits on phylogenies—an approach using the bootstrap. *Evolution* 39:783–791.

Flicek P, Aken BL, Beal K, et al. (59 co-authors). 2008. Ensembl 2008. *Nucleic Acids Res*. 36:D707–D714.

Glenner H, Hansen AJ, Sorensen MV, Ronquist F, Huelsenbeck JP, Willerslev E. 2004. Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach. *Curr Biol*. 14:1644–1649.

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol*. 47:9–17.

Gupta RS. 2001. The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol*. 4:187–202.

Gupta RS. 2006. Molecular signatures (unique proteins and conserved indels) that are specific for the epsilon proteobacteria (Campylobacterales). *BMC Genomics*. 7:167.

Gupta RS, Mok A. 2007. Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. *BMC Microbiol*. 7:106.

Haen KM, Lang BF, Pomponi SA, Lavrov DV. 2007. Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? *Mol Biol Evol*. 24:1518–1527.

Halanych KM. 2004. The new view of animal phylogeny. *Annu Rev Ecol Evol Syst*. 35:229–256.

Hooper JNA, Van Soest RWM. 2006. A new species of *Amphimedon* (Porifera, Demospongiae, Haplosclerida, Niphatidae) from the Capricorn–Bunker Group of Islands, Great Barrier Reef, Australia: target species for the 'sponge genome project'. *Zootaxa*. 1314:31–39.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Res*. 9:868–877.

Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW. 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. *Mol Biol Evol*. 24:1604–1607.

Iseli C, Jongeneel CV, Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*. 138–148.

King N, Westbrook MJ, Young SL, et al. (36 co-authors). 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–788.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*. 24:1380–1383.

Larson A. 1994. The comparison of morphological and molecular data in phylogenetic systematics. *Exs*. 69:371–390.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7(Suppl 1):S4.

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

Lecointre G, Philippe H, Van Le HL, Le Guyader H. 1993. Species sampling has a major impact on phylogenetic inference. *Mol Phylogenet Evol*. 2:205–224.

Lloyd DG, Calder VL. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. *J Evol Biol*. 4:9–21.

Medina M, Collins AG, Silberman JD, Sogin ML. 2001. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proc Natl Acad Sci USA*. 98:9707–9712.

Müller K. 2006. Incorporating information from length-mutational events into phylogenetic analysis. *Mol Phylogenet Evol*. 38: 667–676.

Müller K. 2005. SeqState: primer design and sequence statistics for phylogenetic DNA datasets. *Appl Bioinformatics*. 4:65–69.

Philip GK, Creevey CJ, McInerney JO. 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, Derelle R, Lopez P, et al. (21 co-authors). 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 19:706–712.

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, Laurent J. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev*. 8:616–623.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 35: D61–D65.

Raff RA. 1996. The shape of life: genes, development, and the evolution of animal form. Chicago (IL): University of Chicago Press.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007a. Analysis of rare amino acid replacements supports the coelomata clade. *Mol Biol Evol*. 24:2594–2597.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007b. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol*. 24:1080–1090.

Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*. 15:454–459.

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

Roy SW, Irimia M. 2008. Rare genomic characters do not support Coelomata: RGC_CAMs. *J Mol Evol*. 66:308–315.

Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, Dellaporta SL, Kolokotronis SO, Desalle R. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol*. 7:e20.

Simmons MP, Muller K, Norton AP. 2007. The relative performance of indel-coding methods in simulations. *Mol Phylogenet Evol*. 44:724–740.

Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol*. 49:369–381.

Simmons MP, Ochoterena H, Carr TG. 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. *Syst Biol*. 50:454–462.

Swofford DL. 2003. Paup* phylogenetic analysis using parsimony (*and other methods) version 4.0b10. Sunderland (MA): Sinauer Associates.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Templeton AR. 1983. Phylogenetic ınference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37:221–244.

van Dijk MAM, Paradis E, Catzeflis F, de Jong WW. 1999. The virtues of gaps: xenarthran (Edentate) monophyly supported by a unique deletion in alpha A-crystallin. *Syst Biol*. 48:94–106.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.

Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res*. 14:29–36.

Yandell M, Mungall CJ, Smith C, Prochnik S, Kaminker J, Hartzell G, Lewis S, Rubin G, M. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput Biol*. 2:e15.