Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation

Jie Zheng, Igor B. Rogozin, Eugene V. Koonin*, Teresa M. Przytycka*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Running Head: Intron pattern analysis supports coelomata

Key words: Intron loss; Intron conservation; Phylogenetic analysis; coelomata; ecdysozoa.

* Corresponding authors
E-mail: przytyck@ncbi.nlm.nih.gov or koonin@ncbi.nlm.nih.gov
Abstract

Many intron positions are conserved in varying subsets of eukaryotic genomes and, consequently, comprise a potentially informative class of phylogenetic characters. Roy and Gilbert developed a method of phylogenetic reconstruction using the patterns of intron presence-absence in eukaryotic genes and, applying this method to the analysis of animal phylogeny, obtained support for an Ecdysozoa clade (Roy and Gilbert 2005). The critical assumption in the method was the independence of intron loss in different branches of the phylogenetic tree. Here, this assumption is refuted by showing that the branch-specific intron loss rates are strongly correlated. We show that different tree topologies are obtained, in each case with a significant statistical support, when different subsets of intron positions are analyzed. The analysis of the conserved intron positions supports the Coelomata topology, i.e., a clade comprised of arthropods and chordates, whereas the analysis of more variable intron positions favors the Ecdysozoa topology, i.e., a clade of arthropods and nematodes. We show, however, that the support for Ecdysozoa is fully explained by parallel loss of introns in nematodes and arthropods, a factor that does not contribute to the analysis of the conserved introns. The developed procedure for the identification and analysis of conserved introns and other characters with minimal or no homoplasy is expected to be useful for resolving many hard phylogenetic problems.
Introduction

Traditionally, molecular phylogenetics operates with alignments of homologous nucleotide or protein sequences that are used as the input for phylogenetic tree construction with one or another of the enormous variety of the available methods (Felsenstein 2004). Sequencing of numerous genomes from diverse taxa enabled the extension of phylogenetic analysis to the whole-genome scale. Most often, this involves construction of phylogenetic trees from concatenated alignments of numerous genes or combination of numerous trees for individual gene sets into a supertree, but characters that can be properly denoted as genomic, such as gene composition, gene order, and protein domain combinations, have been employed as well (Snel, Bork, and Huynen 1999; Wolf et al. 2001; Wolf et al. 2002; Snel, Huynen, and Dutilh 2005).

Methodologically, perhaps, the most promising category of genomic characters are rare genomic changes (RGCs) that represent, essentially, the genomic version of shared derived characters (synapomorphies or ‘Hennigian’ markers) (Rokas and Holland 2000; Nei and Kumar 2001; Delsuc, Brinkmann, and Philippe 2005; Boore 2006). As long as homoplasy (parallel, independent evolution of the same character in different lineages) can be ruled out, shared derived characters serve as markers of bona fide clades. Insertion and deletion (gain and loss) of introns in protein-coding genes during the evolution of eukaryotes has been proposed as a promising class of RGCs (Nei and Kumar 2001). Indeed, intron positions in eukaryotic genes appear to be an attractive substrate for phylogenetic analysis because introns are extremely numerous, the positions of many but by no means all introns are conserved even between very distant eukaryotic taxa (Fedorov, Merican, and Gilbert 2002; Rogozin et al. 2003), and independent gain of introns in the same position in different lineages, which would lead to homoplasy,
appears to be rare (Sverdlov et al. 2005). Despite these potential advantages of intron positions as phylogenetic characters, there is a severe problem that, potentially, jeopardizes this approach, namely, the extensive parallel loss of introns in different lineages which leads to gross distortions of phylogenetic trees constructed on the basis of alignments of intron positions (Rogozin et al. 2003). To overcome this problem, Roy and Gilbert devised a statistical technique aimed at distinguishing between alternative phylogenetic hypotheses by comparing patterns of intron conservation (Roy and Gilbert 2005).

Roy and Gilbert applied their method to the same set of alignments of intron positions that was previously analyzed by Rogozin et al. (Rogozin et al. 2003) and addressed one of the most hotly debated, persistent problems in the large-scale animal phylogeny, namely, the controversy surrounding the Coelomata and Ecdysozoa topologies of the phylogenetic tree of animals. The “textbook” tree topology, originally stemming from comparative anatomy, includes a clade of animals that possess a true body cavity (coelomates, such as arthropods and chordates), whereas animals that have a pseudocoelome, such as nematodes, and those without a coelome, such as flatworms, occupy more basal positions in the tree (e.g., (Brusca and Brusca 1990; Raff 1996)). The Coelomata topology appears “natural” from the viewpoint of the straightforward and intuitive concept of the hierarchy of morphological and physiological complexity among animals, which is the main reason why this phylogeny had been accepted since the work of Haeckel (Haeckel 1866). The first molecular phylogenetic analyses of 18S rRNA have supported the Coelomata clade (Field et al. 1988; Turbeville et al. 1991). However, in a seminal 1997 study, Aguinaldo et al. reported a new phylogenetic analysis of 18S rRNAs
from a much larger set of animal species and arrived at an alternative tree topology that clustered arthropods and nematodes in a clade of molting animals termed Ecdysozoa (Aguinaldo et al. 1997).

The Ecdysozoa topology was recovered only when certain, apparently, slowly-evolving species of nematodes were included in the analyzed sample. Accordingly, it has been proposed that the coelomate topology is an artefact caused by long-branch attraction (LBA) (Aguinaldo et al. 1997; Telford and Copley 2005) which is one of the most common artefacts of phylogenetic analysis (Felsenstein 1978; Reyes, Pesole, and Saccone 2000; Philippe et al. 2005). Specifically, the purported LBA has been attributed to the inclusion of fast-evolving species, such as nematodes of the genus *Caenorhabditis*. The ecdysozoan topology received additional support from the results of an independent phylogenetic analysis of 18S RNA (Giribet et al. 2000; Peterson and Eernisse 2001), combined analysis of 18S and 28S rRNA sequences (Mallatt and Winchell 2002), and some protein phylogenies, such as those for Hox proteins (de Rosa et al. 1999). Furthermore, an apparent derived shared character of the Ecdysozoan clade has been identified, a distinct, multimeric form of β-thymosin (Manuel et al. 2000).

Being compatible with the interpretation of molting as a fundamental developmental feature, the ecdysozoan topology was rapidly and nearly universally accepted in the evo-devo community (Adoutte et al. 2000; Valentine and Collins 2000; Collins and Valentine 2001; Telford and Budd 2003). However, phylogenetic analyses of multiple sets of orthologous proteins have reopened the Coelomata-Ecdysozoa conundrum by consistently supporting the Coelomata topology (Mushegian et al. 1998; Blair et al. 2002; Wolf, Rogozin, and Koonin 2004). Both Blair et al. and Wolf et al.
assessed the potential effect of branch length on the tree topology and concluded that the observed support for the Coelomata topology could not stem from LBA (Blair et al. 2002; Wolf, Rogozin, and Koonin 2004). Trees constructed by using non-sequence-based criteria, such as gene content and multidomain protein composition also supported Coelomata (Wolf, Rogozin, and Koonin 2004). Subsequently, the Coelomata topology received further support from several independent phylogenetic studies (Stuart and Berry 2004; Philip, Creevey, and McInerney 2005; Zdobnov et al. 2005; Ciccarelli et al. 2006).

In addition, the status of the multimeric β-thymosin as a derived shared character supporting the Ecdysozoa has been put into doubt as a result of the comparative analysis of recently sequenced genomes (Telford 2004a).

These reports have prompted further re-analyses including large-scale maximum-likelihood phylogenetic analyses of multiple genes from an extended range of animal species (Brinkmann et al. 2005; Dopazo and Dopazo 2005; Philippe, Lartillot, and Brinkmann 2005), putative derived characters, such as shared orthologs and domain combinations (Copley et al. 2004), and patterns of intron conservation in the aforementioned study of Roy and Gilbert (Roy and Gilbert 2005) and, independently, in a similar analysis of Nguyen et al. (Nguyen, Yoshihama, and Kenmochi 2005). All these studies provided support for the Ecdysozoa topology suggesting, once again, that the coelomate topology was an LBA artefact, caused, largely, by inadequate taxon sampling and also, possibly, by the use of over-simplified models of sequence evolution (Brinkmann et al. 2005; Philippe, Lartillot, and Brinkmann 2005; Lartillot, Brinkmann, and Philippe 2007).
Given the multiple reports in support of each of the alternative tree topologies, the Coelomata-Ecdysozoa dilemma is often considered unresolved, and accordingly, the metazoan tree is presented as a multifurcation (Hedges 2002; Telford 2004b; Jones and Blaxter 2005). Very recently, we have re-examined the problem using a new class of RGCs that include lineage-specific replacements of amino acids that are, otherwise, conserved in a broad range of taxa and require two or three nucleotide substitutions; this study provided a strong support for the Coelomata topology (Rogozin et al. 2007). However, in a pattern of controversy that has become characteristic of the Coelomata-Ecdysozoa conundrum, it has been claimed by Irimia et al. that an analysis of the same class of RGCs with a different outgroups supported Ecdysozoa (Irimia et al. 2007).

In an independent line of development, a graph-theoretical method for identifying unstable phylogenetic characters, recently developed by one of us, has been applied to remove, from intron position data, those positions that were found to be prone to multiple intron losses or independent gain. Phylogenetic analysis of the remaining intron positions strongly supported the Coelomata topology (Przytycka 2006; Przytycka 2007).

These findings prompted us to re-examine, in detail, the use of intron position conservation for phylogenetic inference, using animal phylogeny as the test ground. The central assumption of the method developed by Roy et al. is that the probability of retention of a given intron in the given branch can be modeled by a memory-less Markov process (Roy and Gilbert 2005), and the analysis of Nguyen et al. rests on the same assumption (Nguyen, Yoshihama, and Kenmochi 2005). Specifically, the probability of retaining an intron along any tree branch is assumed to depend on the branch but not on the history of the given intron in other branches of the tree. Here we refute this
assumption and demonstrate that an intron is more likely to be conserved along a particular branch if its ortholog is also retained in other branches. Thus, it appears that introns are inherently more stable during evolution in some positions than in others, in an obvious parallel with different sites in proteins. Although the observed dependence of the retention probability on intron conservation in different branches invalidates the analyses of Roy and Gilbert (Roy and Gilbert 2005) and Nguyen et al. (Nguyen, Yoshihama, and Kenmochi 2005), it does not necessarily mean that data on intron presence-absence could not be used to infer phylogenies.

First, we evaluated whether the topology of the resulting tree depends on the level of intron conservation by partitioning the patterns of intron presence-absence into conserved and variable subsets. The analysis of the conserved intron set strongly supports the Coelomata phylogeny, in a direct contradiction with the conclusions of Roy and Gilbert (Roy and Gilbert 2005) and Nguyen et al. (Nguyen, Yoshihama, and Kenmochi 2005), whereas the analysis of the variable introns yields the Ecdysozoa topology. It seems plausible that variable characters (intron positions, in this case) produce an incorrect tree due to long branch attraction. We rigorously show that, in a test of phylogenetic hypotheses with 5 taxa, the outcome is unaffected by parallel losses for characters that are conserved in three or four taxa, but is dramatically biased by parallel losses for characters that are present in only two taxa. In the context of the present study, this means that, if the analysis of the conserved intron set correctly yields the Coelomata topology, then, the Ecdysozoa topology is observed with the variable set due to parallel intron losses (homoplasy). We also present two additional, independent, formal arguments in support of the Coelomata topology.
Materials and Methods

The data set

The data set analyzed here is an extension of the previously described, curated set of conserved eukaryotic genes in which intron positions were mapped onto protein sequence alignments (Rogozin et al. 2003). In addition to the originally analyzed 8 species, namely, Anopheles gambiae (Ag), Arabidopsis thaliana (At), Caenorhabditis elegans (Ce), Drosophila melanogaster (Dm), Homo sapiens (Hs), Plasmodium falciparum (Pf), Saccharomyces cerevisiae (Sc), and Schizosaccharomyces pombe (Sp), two intron-rich fungi, Aspergillus fumigatus (Af) and Cryptococcus neoformans (Cn), and one Apicomplexan, Theileria parva (Tp), were included. The analyzed data set consisted of 585 sets of orthologous genes.

Each intron position corresponds to a binary string, where each bit (1 or 0) corresponds to a species and indicates whether an intron is present (1) or absent (0) at the given position of that species. We call such binary strings “patterns”. To merge multiple species into one group, the patterns of the merged species were replaced by a single pattern that contained 1 if any of the replaced bits was 1, and 0 otherwise. In the present analysis, each pattern is a 5-bit string, including 4 animals and one outgroup (which in some cases consisted of several species). The following three outgroups were considered: Out 1 - Arabidopsis (At), Out 2 - four fungal species (Af, Cn, Sp, Sc), and Out 3 - all non-animal species in the data set (At, Af, Cn, Sp, Sc, Tp, Pf). In the main body of the paper, we present the results for the largest outgroup (Out3) as it provides the most robust data. The results for the other outgroups are presented in the Supplementary Materials.
For brevity, we usually refer to an intron position as an “intron”, and when there is an intron in that position of a species, it is said that the intron occurs in that species. An intron position is considered conserved in two or more species if the intron occurs between two aligning bases in the alignment of the coding sequences. An intron occurs in a group if it occurs in any species of that group. To be informative for phylogenetic analysis, an intron must occur in at least two groups. This intron set was partitioned into conserved and variable subsets. The former subset consists of introns that occur in at least three groups, and the latter subset consists of the introns that occur in only two groups. The numbers of introns in the conserved and variable subsets are shown in Table 1.

**Intron retention rates**

The notations for comparison of retention rates are adopted from the work of Roy and Gilbert (Roy and Gilbert 2005). Let A stand for arthropods (Dm and Ag), D for deuterostomes (Hs), N for nematodes (Ce), and O for outgroup. Then, ANO is the number of introns present in A, N, and O but absent in D, and ADON is the number of introns present in all four taxa. The retention rate is the ratio of introns present in a taxon to the number of introns that are absent (in the analyzed positions). For example, under the Coelomata phylogeny, for introns present in D and O, ADO/DO is the ratio of the number of introns retained along the branch A to the number of introns lost; similarly, ADON/DON is the retention rate in branch A for the introns that are also present in N. The p-values are the probabilities that the two ratios are equal, calculated using Fisher's exact test.

**Dollo parsimony analysis of intron conservation patterns**
Phylogenetic analysis of intron conservation patterns was performed using the Dollo parsimony method (Farris 1977; Rogozin et al. 2005) which was applied either to all introns, or separately to the conserved and variable subsets using the Dolpenny program of the Phylip package (Felsenstein 1996) with default settings. For this analysis, the arthropods were split into two taxa, Dm and Ag, because conserved introns (those that occur in more than two taxa) cannot resolve phylogenies for four taxa.

The statistical significance of the results was assessed using the winning sites test (Prager and Wilson 1988) as follows. For each intron, the minimum numbers of losses in the Ecdysozoa tree \((E)\) and the Coelomata tree \((C)\) was calculated, allowing for one gain only (the assumption of Dollo parsimony). If \(E = C\), the intron is uninformative with respect to the support for one or the other topology and is discarded from the analysis; if \(E > C\), the given intron supports Coelomata, and if \(E < C\), the intron supports Ecdysozoa. Since, as observed from the data, for any intron, \(|E - C| \leq 1\), the numbers of introns that favor each topology reflect the signals of Dollo parsimony. The null hypothesis is that the probability that an intron favors Coelomata or Ecdysozoa is equal to 0.5. Using the binomial distribution, the p-value of a topology was calculated as the probability that at least the observed number of introns favor that topology. It should be noted, for clarity, that introns that are shared by two close sister species only cannot provide support for any hypothesis and are uninformative.

**Results**

**The dependence of intron loss rate in a branch on intron conservation in other branches**
The central assumption of the methods employed by Roy and Gilbert’s (Roy and Gilbert 2005) and Nguyen et al. (Nguyen, Yoshihama, and Kenmochi 2005) is that the probability of retention of a given intron in one branch can be modeled by a memory-less Markov process. Specifically, the probability of retaining an intron along any tree branch is assumed to depend on the branch but not on the retention of the intron in the given position in other branches of the tree. Here, we refute this assumption and demonstrate that an intron in a given position is more likely to be retained along a particular branch if it is also retained in other branches.

We tested the null hypothesis that intron retention rate at any given tree branch is independent on whether or not the intron in the given position is retained in other (independent) branches of the tree. The hypothesis was tested for each of the two alternative topologies of the animal tree, Coelomata and Ecdysozoa. Let X be the common ancestor of Arthropods and Nematodes (for the Ecdysozoa tree) or Arthropods and Deuterostomes (for the Coelomata tree). Consider introns that are present in the outgroup (O) and in at least one child of X. Under the assumption that the introns in the same position are orthologous (no parallel intron gain), such an intron must be present also in the node X but may or may not be present in D (for Ecdysozoa) or N (for Coelomata). We test the null hypothesis that the retention rates \( r_A, r_N \) and \( r_D \) (as appropriate for the corresponding tree topology) do not depend on whether or not the intron was retained in D (for Ecdysozoa) or N (for Coelomata) (Figure 1). If a lineage is represented by more than one species, the presence of the intron in any of these species implies that the intron is present at the root of the lineage. Then, testing the null hypothesis reduces to testing the following equalities:
for Coelomata:
\[
\frac{r_A}{1 - r_A} = \frac{ADO}{DO} = \frac{ADON}{DON}
\]
\[
\frac{r_D}{1 - r_D} = \frac{ADO}{AO} = \frac{ADON}{AON}
\]

for Ecdysozoa:
\[
\frac{r_A}{1 - r_A} = \frac{ANO}{NO} = \frac{ANOD}{NOD}
\]
\[
\frac{r_N}{1 - r_N} = \frac{ANO}{AO} = \frac{ANOD}{AOD}
\]

The results for the largest outgroup (Out 3) are presented in Table 2; compatible results were obtained with the other two outgroups (Supplemental Tables S1 and S2). The null hypothesis consistently fails the Fisher’s exact test, regardless of whether the Coelomata topology or the Ecdysozoa topology is assumed. The difference in intron retention rates depending on the retention in other branches was not only statistically significant but, at least with some outgroups, quite dramatic. For example, for the Ecdysozoa topology and the largest outgroup, Out 3 (all non-animal species), the retention rate of introns present in the nematodes was \(~0.43\), whereas the corresponding retention rate for introns missing in the nematode was only \(~0.15\) (Table 2). Thus, the assumption that the probability of retaining an intron present in an ancestral node is independent on the retention of this intron in other branches of the tree is rejected.

**Parsimony analysis of conserved introns supports the Coelomata topology**

Having shown that the events of intron loss in various branches are strongly dependent, we reasoned that parallel loss that would distort the results of phylogenetic tree reconstruction is expected to be much more common among poorly conserved introns.
than among highly conserved ones. To examine the possible effect of intron conservation on the results of tree construction with the Dollo parsimony method, we analyzed trees for five species from the taxa of interest [Dm and Ag (Anthropods), Ce (Nematodes), Hs (Deuterostomes)] and an outgroup. In this setting, we define variable introns as those that are present in two groups only. Conserved introns are defined as those that are present in at least tree groups. (For this analysis, 5 group are required because, otherwise, the tree would be unresolved; therefore, it was necessary to treat the two Arthropod species as separate groups). The results for various choices of the outgroup are shown in Table 3. We call an intron position informative if the number of losses of the intron in this position differs between the Ecdysozoa and Coelomata tree topologies (Figure 2). The distribution of introns in informative positions provides support for one of the two trees.

We found that conserved introns consistently and highly significantly supported the Coelomata topology, whereas variable introns yielded the Ecdysozoa tree. Given that variable introns substantially outnumber conserved introns, it is not surprising that the tree constructed using all introns was consistent with the Ecdysozoa topology (Table 3). We repeated this analysis for other sets of 5 species each, by replacing human by sea urchin and one of the insects by honey bee, and obtained consistent results (Supplementary Table S3).

**Three additional tests to resolve the Coelomata/Ecdysozoa dilemma**

In the preceding section, we demonstrated that the Dollo parsimony tree constructed using conserved introns is consistent with the Coelomata hypothesis whereas the variable introns (and all introns, given that the variable introns comprise a substantial majority) supported the Ecdysozoa topology. Intuitively, it seems likely that the variable
introns produce an erroneous result due to parallel losses in different branches. Nevertheless, we sought specific, quantitative tests to distinguish between the two topologies. Three independent tests were developed, each based on a specific assumption about the evolutionary model. Since such models inevitably involve some level of simplification with respect to the actual evolutionary scenario, we consider several increasingly more realistic assumptions. A corollary of this test is that the inconsistency between topologies obtained with the conserved and variable (and the full – given that variable introns are more numerous than conserved ones) sets of introns results from parallel intron loss in the variable set. Taken together, the results of the tests described below not only lend strong support to the Coelomata topology but, through the above corollary, prove that the analysis of the unfiltered data leads to an incorrect tree due to parallel intron losses (a form of homoplasy).

**Test 1**

**Assumptions**: The argument is developed under the assumption of Dollo parsimony, i.e., irreversibility of intron loss (no gain in the same position after a loss). However, it also holds for the more general parsimony model where losses are treated as reversible (i.e., a loss and an independent gain of an intron are allowed to occur in the same position) as long as all character changes are weighted equally.

**Argument**: First, we rule out parallel loses (or independent gains) as a possible explanation of the tree topology obtained with conserved introns. Next, we argue that, under the assumption that the tree obtained with conserved introns is correct, the inconsistency between this tree and the tree computed with variable introns can be explained by homoplasy in the latter. It is easy to show that none of the informative
conserved introns could have undergone parallel losses, regardless of the tree topology (Coelomata or Ecdysozoa). For a parallel loss to occur, the intron must be present in the outgroup; otherwise, the parsimonious scenario for any intron will necessarily include gain in a subtree and, at most, a single loss (Fig. 1). Furthermore, for an intron position to be informative, it must contain the intron in one and only one of in the pair of species: *H. sapiens* or *C. elegans* (if the intron is present in both or none, the number of losses would be the same in the Coelomata and Ecdysozoa trees). Consequently, the third species in which the intron is conserved must be one of the insects. Thus, the set of conserved, potentially, informative introns that might be involved in parallel losses is limited to those that are conserved in the outgroup, in exactly one of the species pair *H. sapiens/C.elegans*, and exactly one of the two insect species (Supplemental Table S4 that also contains the corresponding information on the other two ourgroups). However, introns with this pattern of presence-absence have two losses in each of the topologies and thus are not informative. Therefore, none of the informative, conserved introns have parallel losses in any of the two trees, i.e., we proved that, in the case of conserved introns, parallel losses could not bias the result toward the Coelomata or Ecdysozoa tree. Thus, we have eliminated the possibility that the result is biased in any direction by homoplasy.

The second part of the argument deals with the properties of informative variable introns. As in the case of conserved introns, all informative variable introns must be present in exactly one of the species pair *H. sapiens/C. elegans* (otherwise, they are not informative). Since a variable intron, by definition, is only shared by two lineages, the other lineage has to be either the outgroup or one of the insects. Thus, there are four sets
of informative variable introns. Using the data for the largest, all non-animal outgroup (Out3), the sets and the corresponding numbers are as follows (see Supplemental table S5 that also contains the corresponding information on the other two outgroups)

1. Introns conserved in human and the outgroup (let \( m_{HO} \) be the number of such introns, \( m_{HO} = 711 \)) – one loss in the Ecdysozoa tree, two parallel losses (\( C. elegans \) and one insect) in the Coelomata tree; supports Ecdysozoa
2. Introns conserved in \( C. elegans \) and the outgroup (let \( m_{CO} \) be the number of such introns; \( m_{CO} = 85 \)) – one loss in the Coelomata tree, two parallel losses (\( H. sapiens \) and the common ancestor of the insects) in the Ecdysozoa tree; supports Coelomata
3. Introns conserved in human and one insect (let \( m_{HI} \) be the number of such introns; \( m_{HI} = 168 \)) - one loss in the Coelomata tree, two parallel losses in the Ecdysozoa tree (the second insect and \( C. elegans \)); supports Coelomata
4. Introns conserved in \( C. elegans \) and one insect (let \( m_{CI} \) be the number of such introns; \( m_{CI} = 26 \)) - one loss in the Ecdysozoa tree and two parallel losses (human and one insect) in the Coelomata tree; supports Ecdysozoa.

Thus, assuming that the Coelomata topology is correct, the support for the Ecdysozoa topology obtained in the analysis of the variable introns results from the pattern of intron distribution in set 1 that is, under this assumption, explained by parallel intron losses (an analog of LBA) in \( C. elegans \) and insects.

This argument holds for any set of five taxa and any set of characters as long as the goal is to differentiate between two pre-defined tree topologies, such as Coelomata and Ecdysozoa. Thus, this approach resolves the discrepancy between the results.
obtained for a conserved set of characters and a variable set of characters in favor of the result obtained with the conserved characters – as long, of course, as the number of such characters is sufficient to obtain statistically significant results.

**Impact of assumptions and simplifications:** The argument holds independent of the assumption of character loss irreversibility (Dollo parsimony). However, it relies on the assumption that all changes have the same cost. This is an oversimplification, and a more realistic scenario should account for the different costs of intron loss in different branches. The second test we developed takes this into consideration.

**Test 2**

**Assumptions:** Intron loss is assumed to be irreversible. We further assume that the cost of intron loss in the human branch and the *C. elegans* branch are different. (We can also associate weights with intron loss in other species but this is irrelevant for the argument). Let $d_H$ and $d_C$ be the costs of intron loss in the human branch and the *C. elegans* branch, respectively. Then, the phylogeny is constructed under a variant of the Dollo parsimony model where each loss is scored according to its assigned weight. Because, under parsimony, no *a priori* knowledge of the lengths of the internal edges is assumed, the costs of intron loss are assumed to be equal for all internal branches. Additionally, it is assumed that the ratio of the costs of loss between the variable and conserved sets of introns is approximately the same for the human and *C. elegans* branches. Since it is known that introns are lost on a massive scale in *C. elegans* but are highly conserved in humans (Fedorov et al. 2003; Roy and Penny 2006; Carmel et al. 2007), we assume $d_H/d_C > 1$. 
Argument: First, we show that the value of $d_H/d_C$ affects the result of the Dollo parsimony reconstruction. Depending on this ratio, the conserved and variable intron sets could, potentially, produce different tree topologies. We addressed the following question: is it possible to select the ratio $d_H/d_C$ such that the trees for variable and conserved introns agree, and if so, what is the topology of the “agreement tree”? Note that, for each of the two sets of introns (the conserved and the variable ones), there exists an equilibrium value $d_H/d_C$ where the Coelomata tree and the Ecdysozoa tree have the same cost (Fig. 2). These equilibrium values can be computed directly from the frequencies of intron patterns in both sets (Tables S4 and S5 in the Supplementary Material) as described below. We show that there exists an interval of $d_H/d_C$ values where the tree constructed from the two sets of introns is the same, and this common tree has the Coelomata topology.

First, we compute the equilibrium value for the variable introns. Let $m_{HI}$, $m_{CI}$, $m_{OH}$, and $m_{OI}$ be the number of variable introns in each of the four groups as defined for Test 1. Let us note that:

1. Introns that are conserved in human and the outgroup have one loss in the Ecdysozoa tree. In the Coelomata tree, these introns have an additional loss in the \textit{C. elegans} branch. Thus, the extra cost to the Coelemata tree contributed by this set of introns is $m_{IO}d_C$.

2. Introns that are conserved in \textit{C. elegans} and the outgroup have one loss in the Coelomata tree. In the Ecdysozoa tree, they have an additional loss in the human branch. Thus, the extra cost to the Ecdysozoa tree from this set of introns is $m_{CO}d_H$. 
3. Introns that are conserved in human and one of the insects have one loss in the Coelomata tree. In the Ecdysozoa tree, these introns have an additional loss in the \textit{C. elegans} branch. Thus, the extra cost to the Ecdysozoa tree from this group of introns is $m_{HI}d_C$.

4. Introns conserved in \textit{C. elegans} and one of the insects, have one loss in the Ecdysozoa tree. In the Coelomata tree, they have an additional loss in the human branch. Thus, the extra cost to the Coelomata tree from this group of introns is $m_{CI}d_H$.

The equilibrium value of the $d_H/d_C$ ratio is obtained from the equation that equalizes these extra costs:

$$m_{HI}d_C + m_{OC}d_H = m_{CT}d_H + m_{OH}d_C.$$  

Substituting specific numbers (Supplementary Table S5), we obtain:

$$\frac{d_H}{d_C} = \frac{m_{HO} - m_{HI}}{m_{CO} - m_{CI}} = 9.3.$$  

If $d_H/d_C > 9.3$, the resulting tree has the Coelomata topology, and otherwise, the tree has the Ecdysozoa topology.

Consider now the set of conserved introns. We showed that the only informative introns among the conserved ones are those that are shared by both insects and either human or \textit{C. elegans} (see Test 1 above). Let $n_{HI}$ be the number of introns that are conserved in human and both insects. These introns have no losses in the Coelomata tree and one loss (in the \textit{C. elegans} branch) in the Ecdysozoa tree. Similarly, let $n_{CI}$ be the number of conserved introns with the intron present in \textit{C.elegans} and both insects. For
this set of introns, there are no losses in the Ecdysozoa tree and one loss (in the human branch) in the Coelomata tree. Thus, the equilibrium ratio is obtained from:

\[ n_{HI}d_C = n_{CI}d_H \]

And, substituting the actual counts \( n_{HI} \) and \( n_{CI} \) (Supplemental Table S4) in the conserved intron set, we get:

\[ \frac{d_H}{d_C} = \frac{n_{HI}}{n_{CI}} = 11.3 \]

Thus, if \( d_{HI}/d_C < 11.3 \), then the tree for the conserved introns has the Coelomata topology whereas, for \( d_{HI}/d_C > 11.3 \), the tree has the Ecdysozoa topology.

The only interval of the \( d_{HI}/d_C \) ratio, for which both sets of introns (conserved and variable) produce the same tree, is 9.3 < \( d_{HI}/d_C \) < 11.3, and in this interval, the agreement tree has the Coelomata topology (Fig. 3)

Possible shortcomings of the assumptions. In this argument, the cost of intron loss varies between species but the cost of intron loss in all internal branches is assumed to be the same. This oversimplification is removed in the next test.

Test 3

Assumptions: For this test, intron loss is assumed to be irreversible but there are no constraints on the relation between the retention rates in different tree branches. It is, however, assumed that the retention rates, i.e., the ratio of the number of retained introns to the number of introns that were lost along a given branch, are higher for conserved than for variable introns.
**Argument:** We tested whether the latter assumption was violated under either of the two compared animal tree topologies. The retention rates were computed, under the Dollo parsimony assumption, for conserved and variable introns under the Coelomata and Ectysozoa topologies (Supplementary Material, Table S6). For the Coelomata topology, in all branches of the tree, the variable introns show a lower than or, approximately, the same retention rate as the conserved introns. By contrast, under the Ecdysozoa topology, the retention rate along the *C. elegans* branch was significantly greater for the variable introns than for the conserved introns ($p < 0.02$, one-sided Fisher test). This is the only significant deviation from the expectation regarding the retention rates that we observed, and it is seen in the Ecdysozoa tree. Thus, the results of this test appear to be best compatible with the Coelomata topology.

4. **Discussion and Conclusions**

Intron positions seem to be attractive candidates for the role of RGCs because eukaryotic genes contain numerous introns, thus, providing for statistically powerful phylogenetic tests and also because parallel gains of introns appear to be rare (Sverdlov et al. 2005). However, parallel intron losses in the same position are much more common and complicate phylogenetic analysis through the attraction of branches with high intron loss rates (a version of LBA). Roy and Gilbert (Roy and Gilbert 2005) as well as Nguyen et al (Nguyen, Yoshihama, and Kenmochi 2005) developed methods of phylogenetic reconstruction that overcame the problem of parallel losses but only under the assumption of independence of the loss rates in different branches of the tree. In both studies, the application of these methods to the animal phylogeny supported the Ecdysozoa clade.
Here, we show that the independence assumption is invalid, i.e., the intron loss rates in different branches are strongly correlated. The outcome of phylogenetic analysis critically depends on the subset of intron positions that are used as the input. We show that, when exactly 5 taxa are used for phylogenetic analysis, for the introns that are conserved in three or four taxa, there are no parallel losses in informative positions, so the correct phylogeny is recovered so long as the assumption of the irreversibility of intron loss holds, at least, approximately, and the number of informative positions is sufficient to make the analysis statistically valid. In the specific case of animal phylogeny examined here, the analysis of such conserved introns strongly supports the Coelomata topology. By contrast, the analysis of variable introns (those represented in two species only) supports the Ecdysozoa topology, and because there are many more variable introns than conserved ones, the Ecdysozoan topology is recovered also when the entire set of introns is analyzed, in agreement with the observations of Roy and Gilbert (Roy and Gilbert 2005) and Ngueyn et al (Nguyen, Yoshihama, and Kenmochi 2005). However, we showed that, if the topology obtained with the set of conserved introns (Coelomata) is valid, the recovery of the alternative topology (Ecdysozoa) is explained by parallel losses, in this case, in nematodes and insects.

With the results presented here, it appears that all tested RGCs including protein domain combinations (Wolf, Rogozin, and Koonin 2004), two-substitution replacements of highly conserved amino acids (Rogozin et al. 2007), and now intron positions support the Coelomata topology of the animal tree and reject the Ecdysozoa topology. However, some alternative analyses of RGCs (Copley et al. 2004; Irimia et al. 2007) and several sequence-based phylogenetic studies employing extensive taxon sampling and
sophisticated models of sequence evolution applied to sets of slowly evolving genes (Brinkmann et al. 2005; Philippe, Lartillot, and Brinkmann 2005; Lartillot, Brinkmann, and Philippe 2007) support the Ecdysozoa and suggest that the Coelomata topology is an LBA artefact. This interpretation does not apply to the results obtained here with the set of conserved introns; moreover, we specifically show that, when intron positions are used as phylogenetic characters, exactly the opposite is true, i.e., it is the Ecdysozoa topology that results from LBA (extensive parallel loss of introns in arthropods and nematodes). The reasons behind the discrepancy remain to be investigated in full. However, it should be noted that the analogy between slowly evolving genes and conserved introns employed here as phylogenetic characters is limited at best. As shown here, the analysis of conserved introns in sets of 5 lineages actually eliminates homoplasy (parallel intron loss) whereas the use of sets of slowly evolving genes (or slowly evolving alignment positions) only have the potential to reduce homoplasy, and to extent that is hard to measure precisely. Of course, the method employed here and the additional tests we devised are not free of their own assumptions which are explicitly stated above. The definitive resolution of the Coelomata-Ecdysozoa dilemma will require detailed analyses of many more genomes from different branches of animals and reconciliation of the results obtained with various types of RGCs with the results of sequence-based phylogenetics.

The phylogenetic methodology described here, in principle, can be applied not only to introns but to any binary characters that meet the assumption of the irreversibility of losses that is required for the use of Dollo parsimony. Previously, attempts have been made to increase the accuracy of sequence-based phylogenies by limiting the analysis to
slowly evolving positions in multiple alignments of protein and rRNA sequences, on the premise that such positions are the ones that are least prone to homoplasy (Brinkmann and Philippe 1999; Philippe, Germot, and Moreira 2000; Brochier and Philippe 2002). However, in this case, homoplasy could only be reduced by an uncertain amount, and there was the inevitable trade-off between the selection of increasingly conserved (presumably, increasingly homoplasy-free) positions and the loss of statistical power. The latter issue is pertinent also for the method described here but, at least, in the present case study, the number of conserved introns, although a minority, was amply sufficient to discriminate between the two competing hypotheses. Moreover, applying this method to 5 taxa and using Dollo parsimony allows one not only to reduce but, actually, to demonstrably eliminate a certain type of homoplasy. Therefore, the method described here is expected to be useful for resolving many hard phylogenetic problems.

Acknowledgements
The authors would like to thank Yuri I. Wolf for valuable discussions. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.
References


Figure legends

Figure 1. Testing the independence of intron loss rates in different branches.
For the Ecdysozoa and Coelomata tree topologies, the null hypothesis is that the intron retention rates along the edges descending from the node X (r_A and either r_N or r_D, depending on the topology) is independent on whether the intron in the given position is also retained in D (for Ecdysozoa) or in N (for Coelomata).

Figure 2. Phylogenetically informative patterns of intron presence-absence.
Top: An informative pattern for a variable intron, with a smaller number of losses in the Ecdysozoa topology compared to the Coelomata topology. Bottom: An informative pattern for a conserved intron, with a smaller number of losses in the Coelomata topology compared to the Ecdysozoa topology. The thick red bars represent intron losses.

Figure 3. The dependence of the tree topology on the d_H/d_C ratio. For the variable introns d_H/d_C = 1 results in the Ecdysozoa tree and changes to the Coelomata tree for d_H/d_C> 9.3. For the conserved introns, d_H/d_C = 1 results in the Coelomata tree and changes to the Ecdysozoa tree for d_H/d_C> 11.3. For the interval 9.3< d_H/d_C <11.3, the trees obtained from both sets agree on the Coelomata topology.
Table 1. The conserved and variable introns in the analyzed gene set, depending on the outgroup.

<table>
<thead>
<tr>
<th>Outgroup</th>
<th>Total number of introns</th>
<th>Variable introns</th>
<th>Conserved introns</th>
<th>Number of genes containing at least one conserved intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out 1</td>
<td>1372</td>
<td>939</td>
<td>433</td>
<td>216</td>
</tr>
<tr>
<td>Out 2</td>
<td>1394</td>
<td>925</td>
<td>469</td>
<td>212</td>
</tr>
<tr>
<td>Out 3</td>
<td>1745</td>
<td>1203</td>
<td>542</td>
<td>242</td>
</tr>
</tbody>
</table>
Table 2. Testing the dependence of intron retention on conservation in other branches.

<table>
<thead>
<tr>
<th></th>
<th>Ecdysozoa</th>
<th>Coelomata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>Non-conserved in D</td>
<td>Non-conserved in N</td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>$r_A / (1 - r_A)$</td>
<td>11/85</td>
<td>131/711</td>
</tr>
<tr>
<td>$r_A / r_D$</td>
<td>0.73</td>
<td>0.9</td>
</tr>
</tbody>
</table>

The numbers in the table are for the Out3 outgroup.
Table 3. Dollo parsimony trees constructed for variable and conserved introns.

<table>
<thead>
<tr>
<th>Outgroup</th>
<th>All introns conserved in at least 2 groups</th>
<th>Introns conserved in exactly 2 groups only</th>
<th>Introns conserved in at least 3 groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ecdysozoa 429/695 p=3.36e-10</td>
<td>Ecdysozoa 419/598 p=1.88e-23</td>
<td>Coelomata 87/97 p=8.94e-17</td>
</tr>
<tr>
<td>Out 1</td>
<td>Ecdysozoa 456/698 p=2.22e-16</td>
<td></td>
<td>Coelomata 81/88 p=2.24e-17</td>
</tr>
<tr>
<td>Out 2</td>
<td></td>
<td></td>
<td>Coelomata 68/74 p=1.07e-14</td>
</tr>
<tr>
<td>Out 3</td>
<td>Ecdysozoa 737/990 p=9.80e-56</td>
<td>Ecdysozoa 731/916 p=1.11e-77</td>
<td></td>
</tr>
</tbody>
</table>

The ratio of the number of informative introns that support a given topology to the total number of informative introns is indicated. The p-value was computed using the binomial test as described in the Methods section.
Figures

Figure 1.
Figure 2.
Figure 3.