Supplementary Figures for

Near Intron Pairs and the Metazoan Tree.

Jörg Lehmann∗, Peter F. Stadler, and Veiko Krauss

∗joe@bioinf.uni-leipzig.de
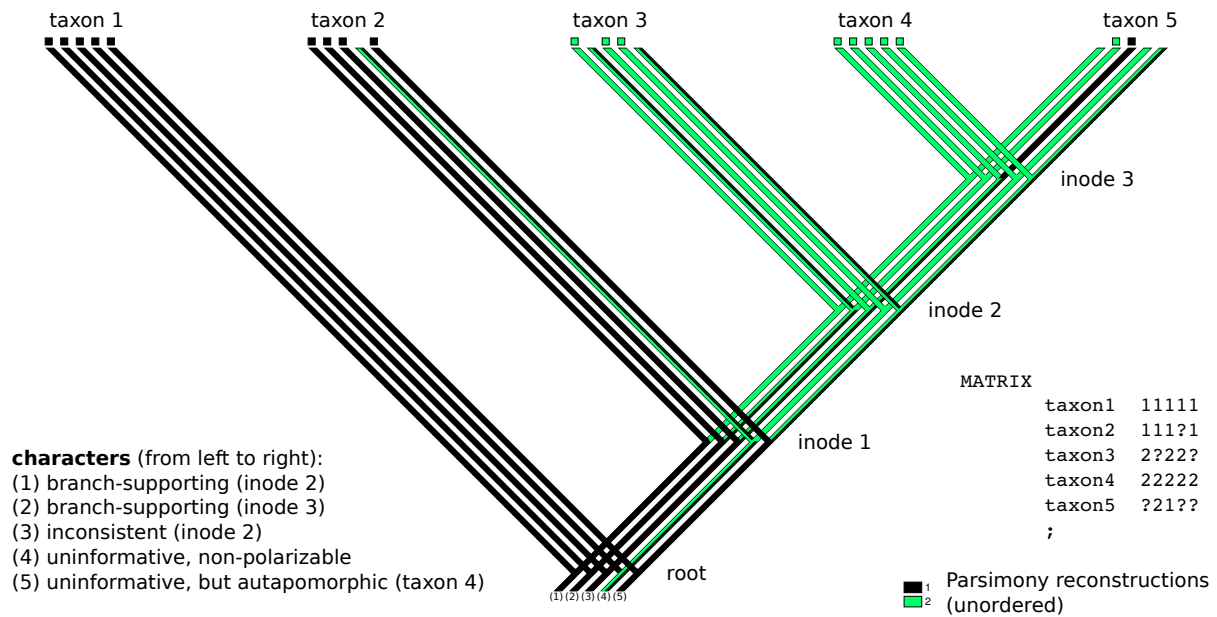
taxon 1　　taxon 2　　taxon 3　　taxon 4　　taxon 5

inode 3

inode 2

MATRIX
```
        taxon1  11111
        taxon2  111?1
        taxon3  2?22?
        taxon4  22222
        taxon5  ?21??
        ;
```

inode 1

characters (from left to right):
(1) branch-supporting (inode 2)
(2) branch-supporting (inode 3)
(3) inconsistent (inode 2)
(4) uninformative, non-polarizable
(5) uninformative, but autapomorphic (taxon 4)

root

(1)(2)(3)(4)(5)

■ 1 Parsimony reconstructions
■ 2 (unordered)

Figure S1: Illustration of the terminology using an artificial example comprising 5 taxa and 5 near intron pair (NIP) characters. For simplicity taxon 1 has character state '1' at all sites (i.e., intron presence at the upstream position). Characters (1)-(3) are parsimony-informative. Character (5) is counted as autapomorphy for taxon 4 as this NIP could be polarized. A character is polarizable w.r.t. a certain tree topology (rooted by a known outgroup), if the ancestral state can be inferred parsimoniously. For this, we require that the character can be mapped onto a specific internal (or external in the case of autapomorphies) branch marking the transition from ancestral to derived state, i.e., at least two terminal taxa that are outside of the derived clade and whose last common ancestor is also the ancestor of the derived clade must have the ancestral state, and the derived state is exclusively observed within the derived clade. Note that these ancestral-state terminal taxa do not necessarily have to include any outgroup taxa used to root the tree. In our setting, a polarized parsimony-informative character is counted as supporting a certain internal node or branch leading to that node (synapomorphy) if taxa with the derived state are observed at least once in every single child clade of that node (exception: multifurcations). In case of additional ancestral-state taxa among the derived subclade, we denote the character as inconsistent w.r.t. the branch leading to the otherwise derived clade. Thus, a branch can have (supporting but) inconsistent character distributions only if at least one of its child nodes is a non-terminal taxon.
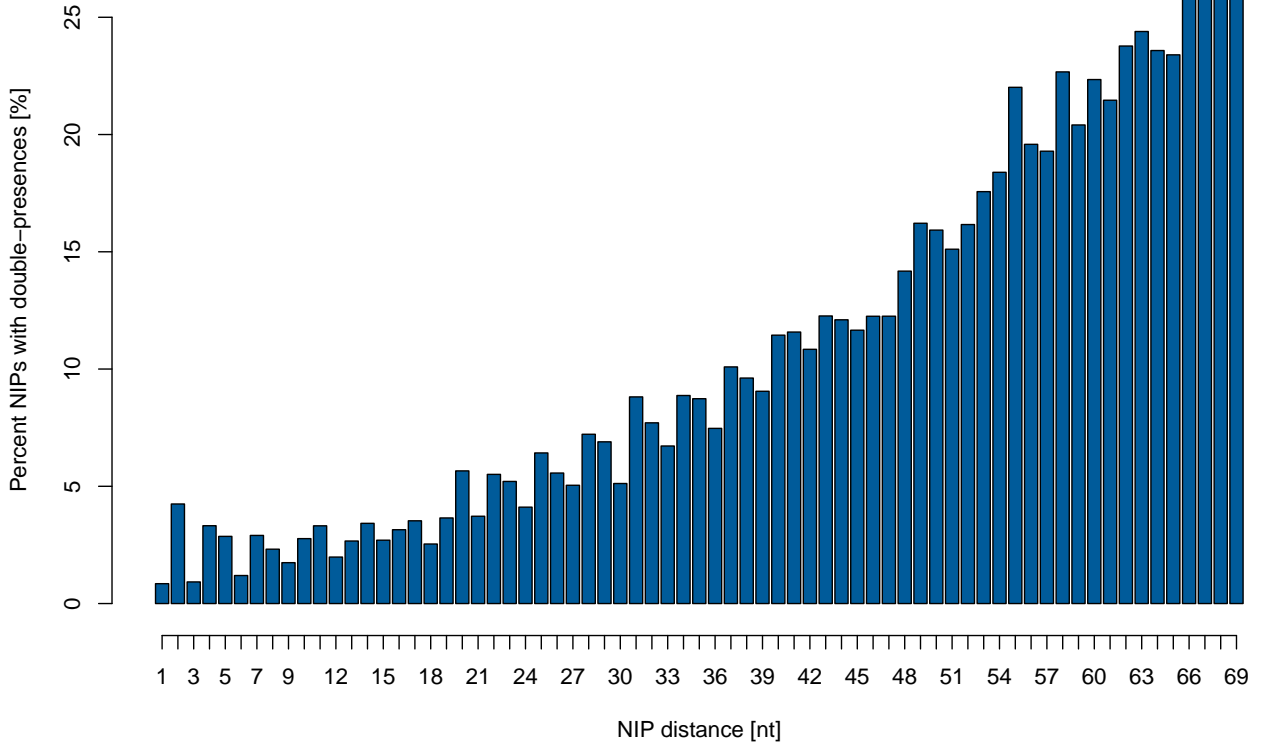
Figure S2: Proportion of NIPs with double-presences (i.e., both intron positions present in a taxon) as a function of NIP distances (d). Data are based on an extended NIP dataset ($d < 70$). For $d < 37$, this fraction is smaller than 9% for each NIP distance. The average proportion of excluded NIPs from the corresponding dataset with $d < 32$ is 3.4%, in comparison to 9.4% for the extended dataset.
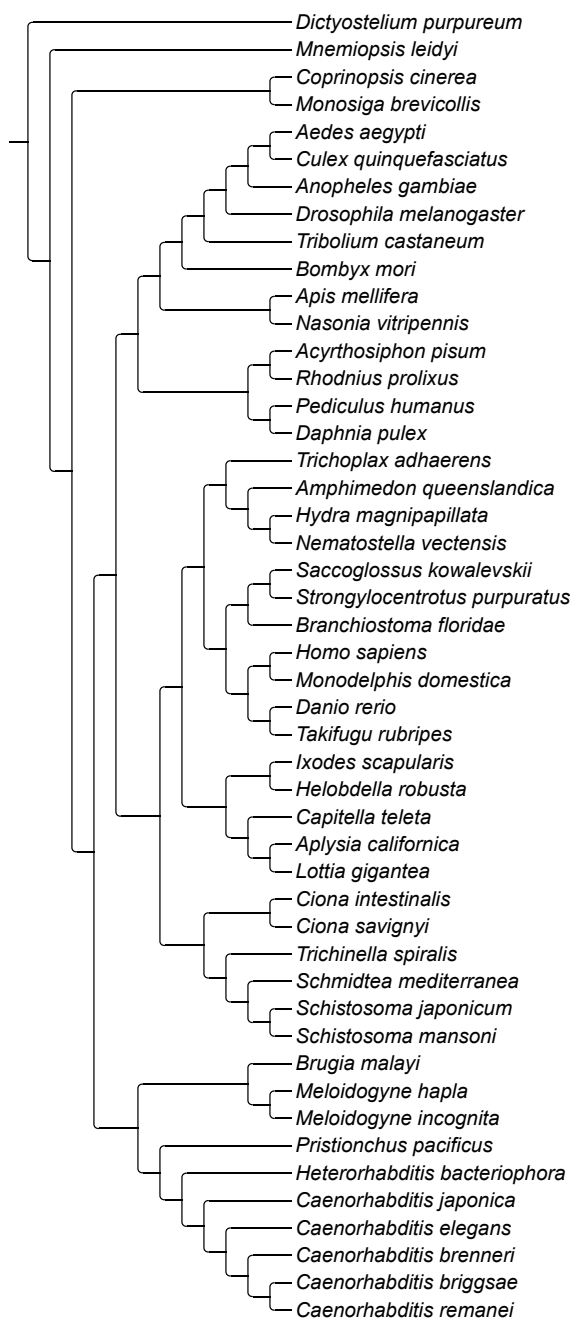
Figure S3: MP tree topology obtained with extended parsimony approach (homoplasy penalty 10) using all 73,593 NIPs. Heuristic search was done with sequential addition (fix order), and subsequent NNI rearrangements.
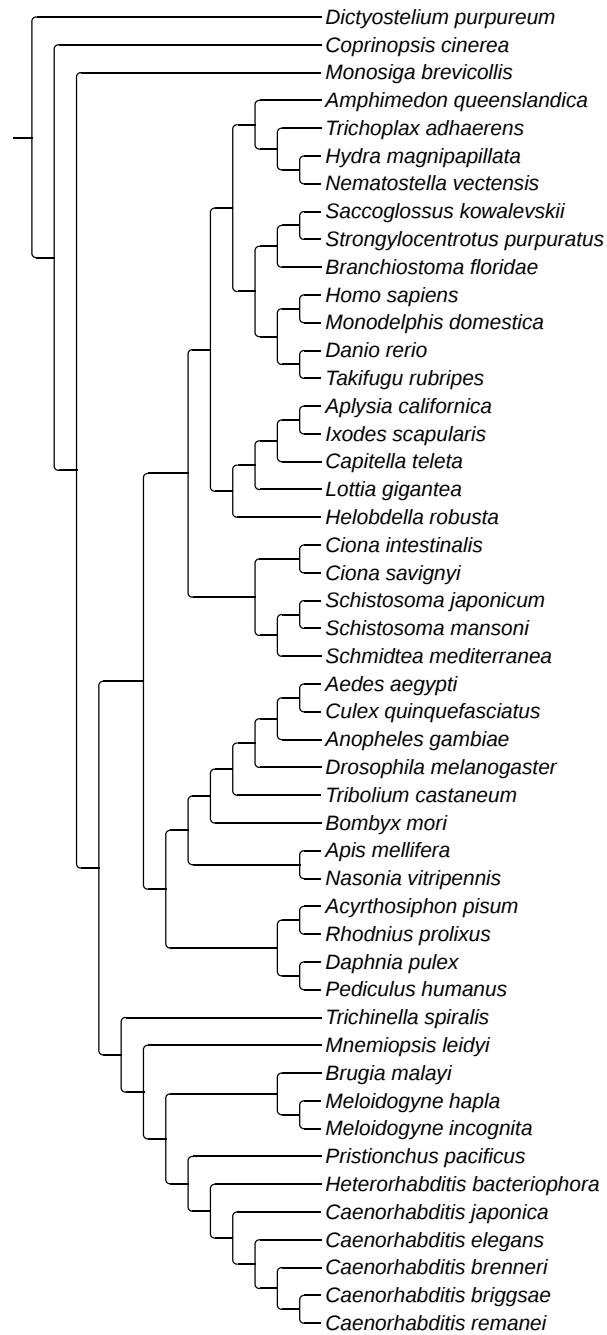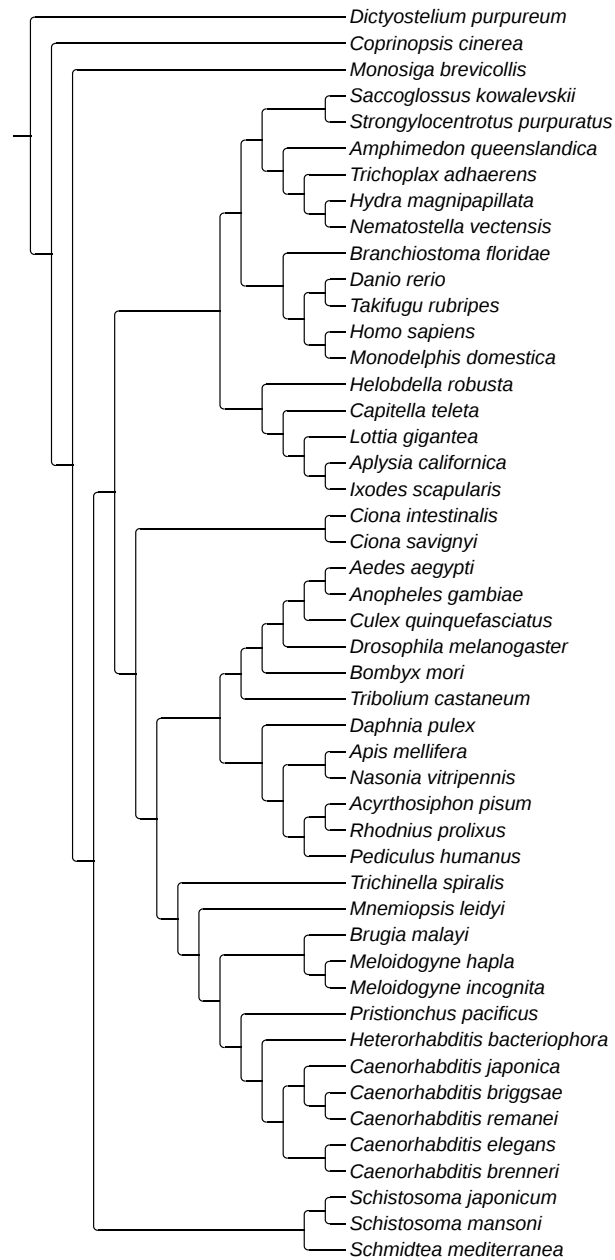
Figure S4: MP tree topology obtained with extended parsimony approach (homoplasy penalty 100) using all 73,593 NIPs. Heuristic search was done with sequential addition (fix order), and subsequent NNI rearrangements.

Figure S5: MP tree topology obtained with an extended parsimony approach (homoplasy penalty 1000) using all 73,593 NIPs. Heuristic search was done with sequential addition (fix order), and subsequent NNI rearrangements.
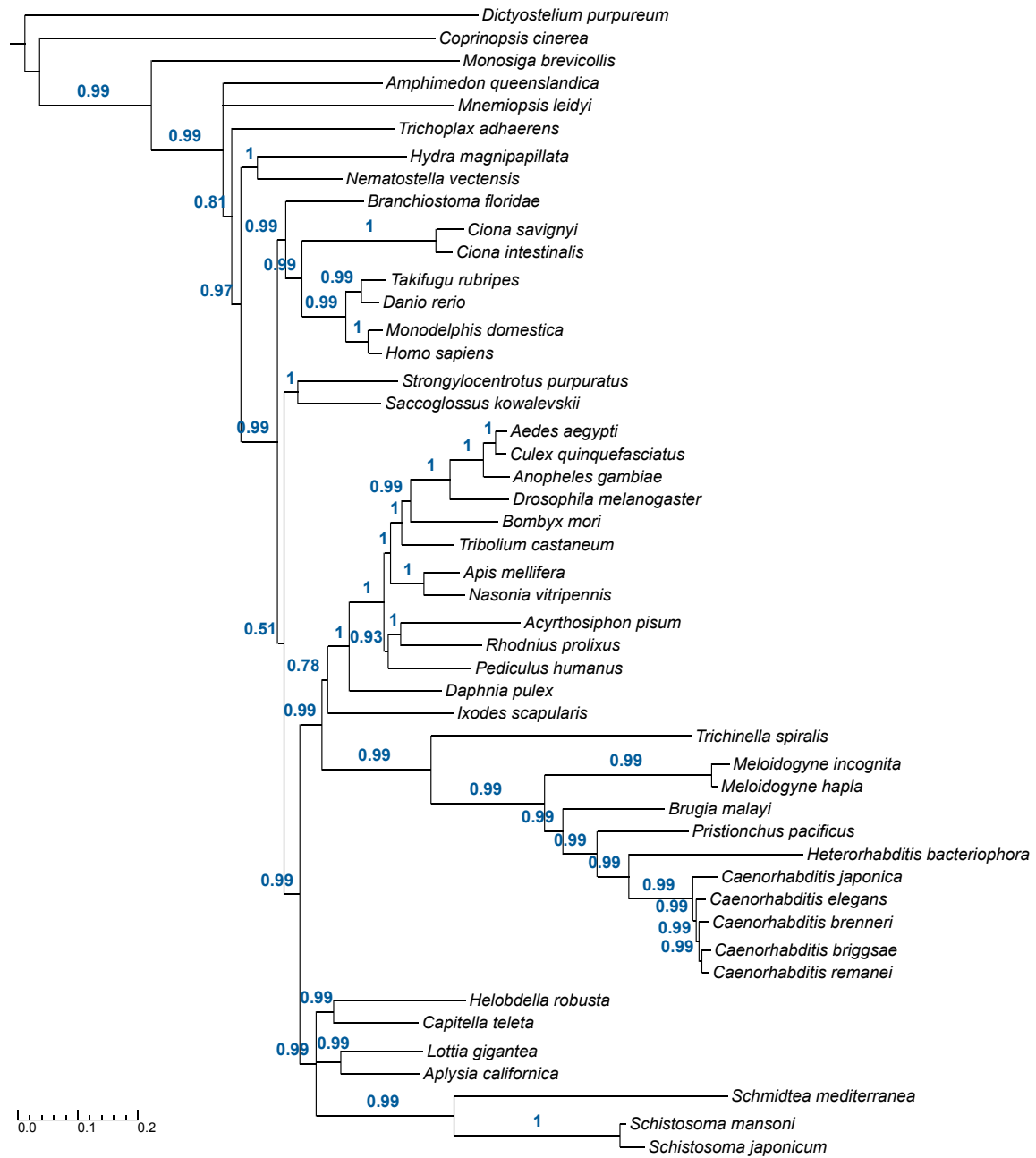
Figure S6: BI consensus tree as obtained from conserved amino acid alignment columns and PhyloBayes. Posterior probability values are given above the corresponding branches.
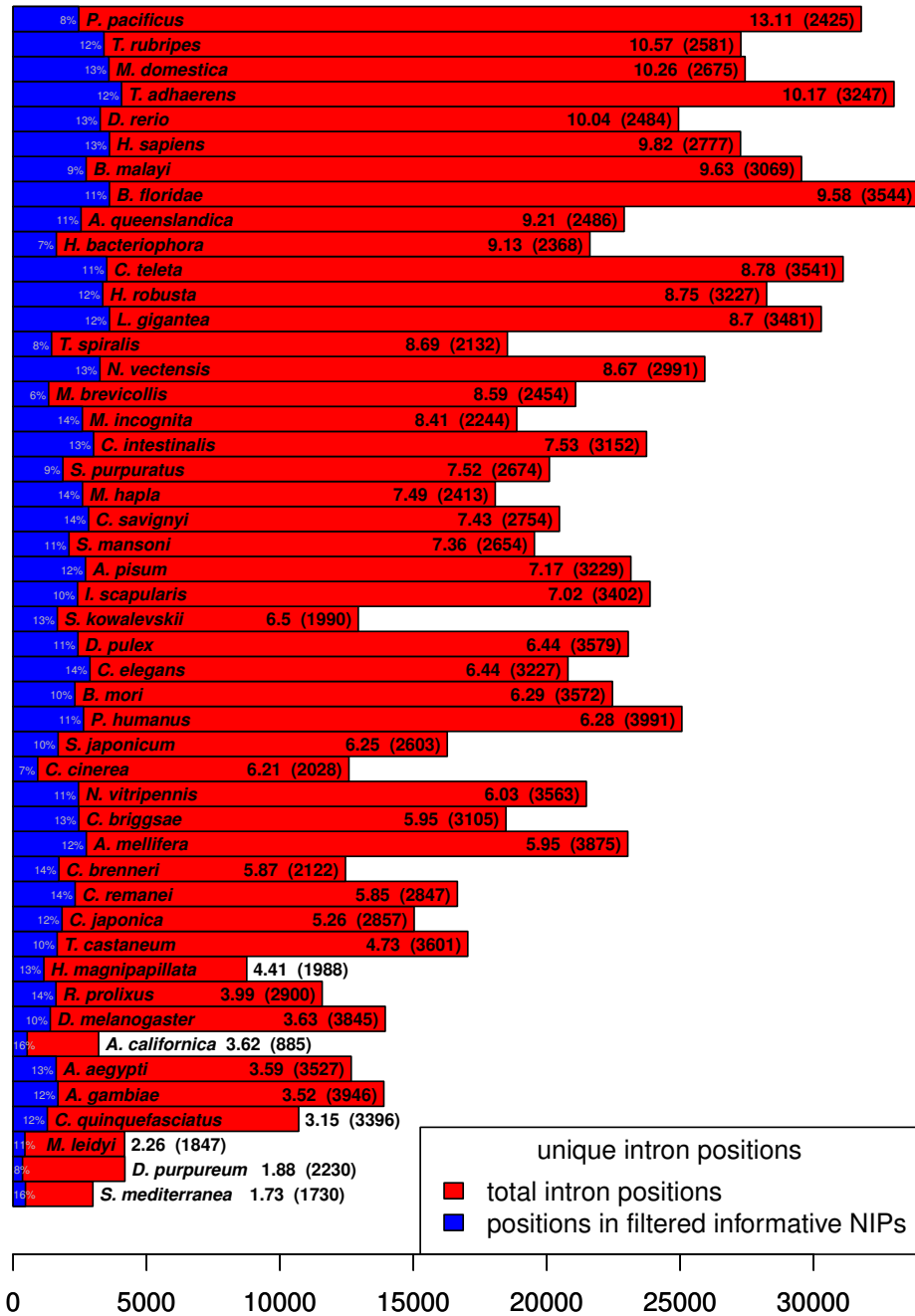
Figure S7: Distribution of total intron numbers (red) and introns involved in parsimony-informative NIPs across all taxa. Taxa are sorted according to the overall intron density (average number of introns per gene) from the full ortholog input dataset (4,405), which is displayed together with the number of ortholog datasets the taxon contributes to (in parentheses). Small total numbers of introns (and consequently few intron positions participating in NIPs) can be explained by a general low intron density within a taxon, and/or a rather low coverage within the ortholog input dataset. Additionally, for each taxon, the fraction of its introns contributing to informative NIPs is specified.
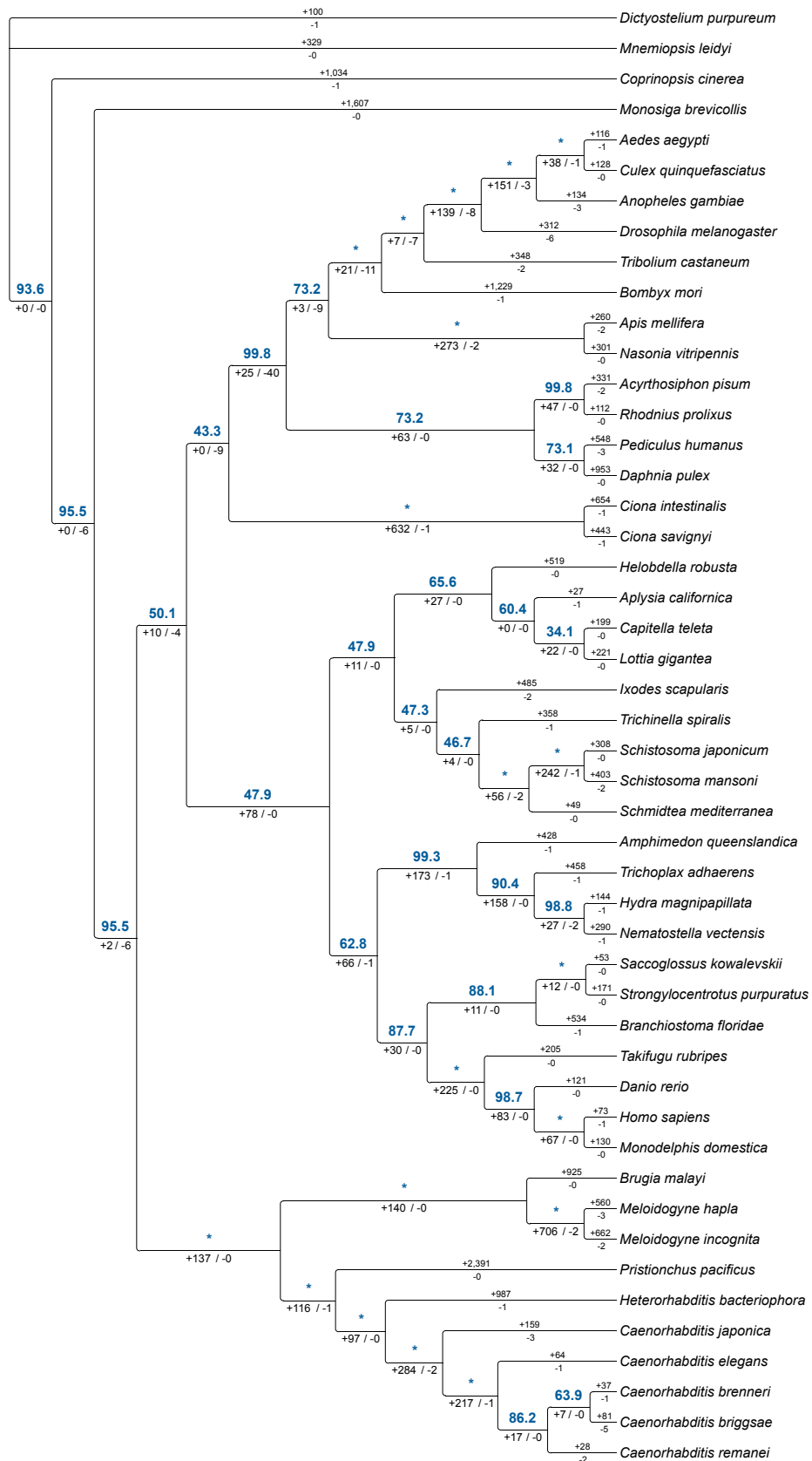
Figure S8: MP strict consensus tree from conserved intron positions (intron presence/absence with Dollo-parsimony). From all 4,405 alignments, the conserved intron positions (gap-free amino acid alignment around intron position (window of 3 aa) and conservation score cutoff 0.5) were collected (44,235 characters) and utilized in heuristic tree searches with Dollo parsimony as character type (16,909 parsimony-informative characters). Bootstrap support (in percent) are given above the corresponding internal branches (in blue), a star indicates 100% support. Additionally, for peach internal branch, the numbers of synapomorphic distributions are displayed, separately for intron gains (+) and losses (-). At the terminal branches, the number of exclusive (autapomorphic) intron gains and losses is displayed. (strict consensus tree length=57,070, CI=0.296, RI=0.878)
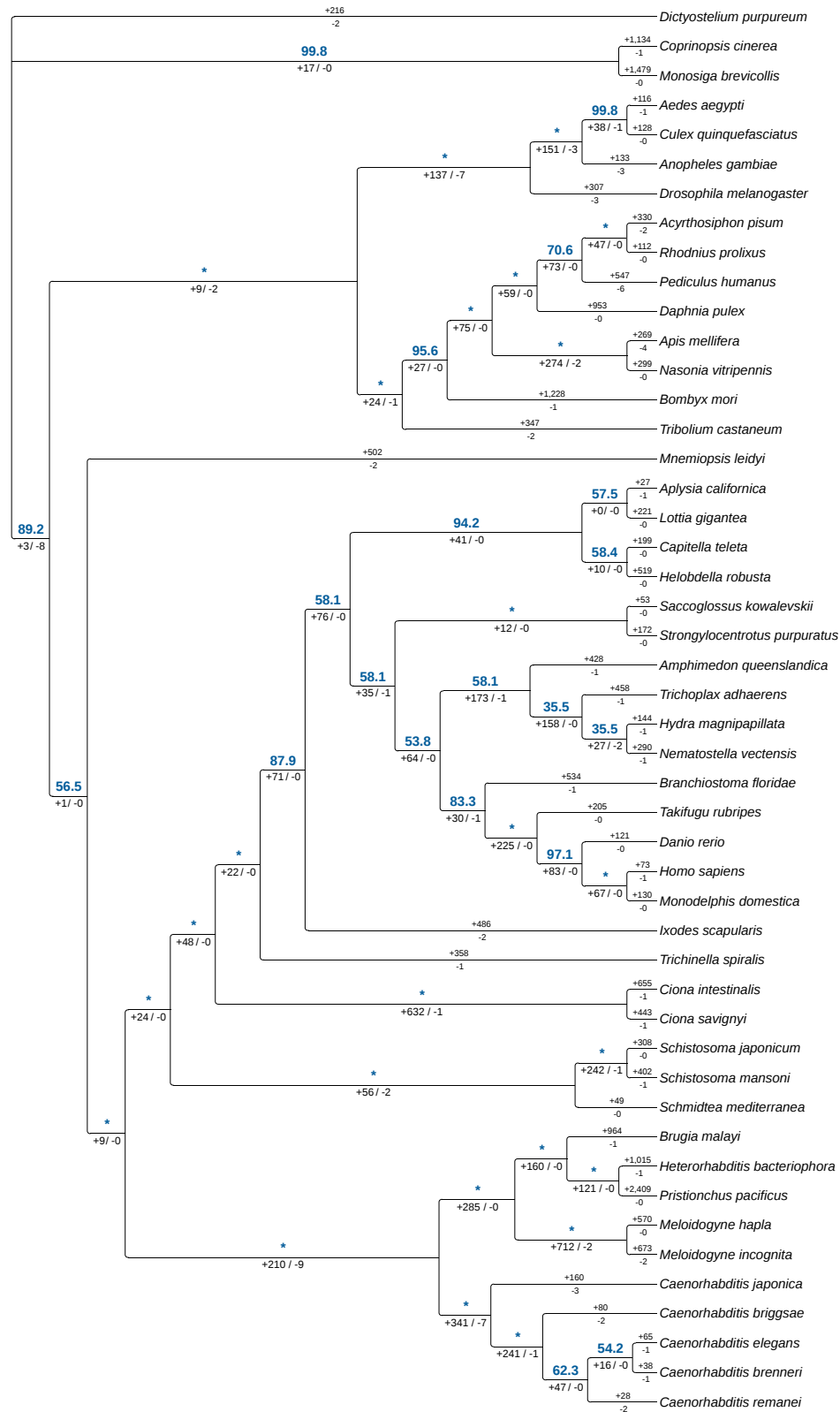
Figure S9: MP strict consensus tree from conserved intron positions (c.f. fig. S8) using Wagner parsimony (16,357 parsimony-informative characters, unordered, equal weights). Bootstrap percentages are given above the corresponding internal branches (in blue), a star indicates 100% support. Additionally for each internal branch, the numbers of synapomorphic distributions are displayed, separately for intron gains (+) and losses (-). At the terminal branches, the number of exclusive (autapomorphic) intron gains and losses is displayed. (strict consensus tree length=40,988, CI=0.399, RI=0.686)
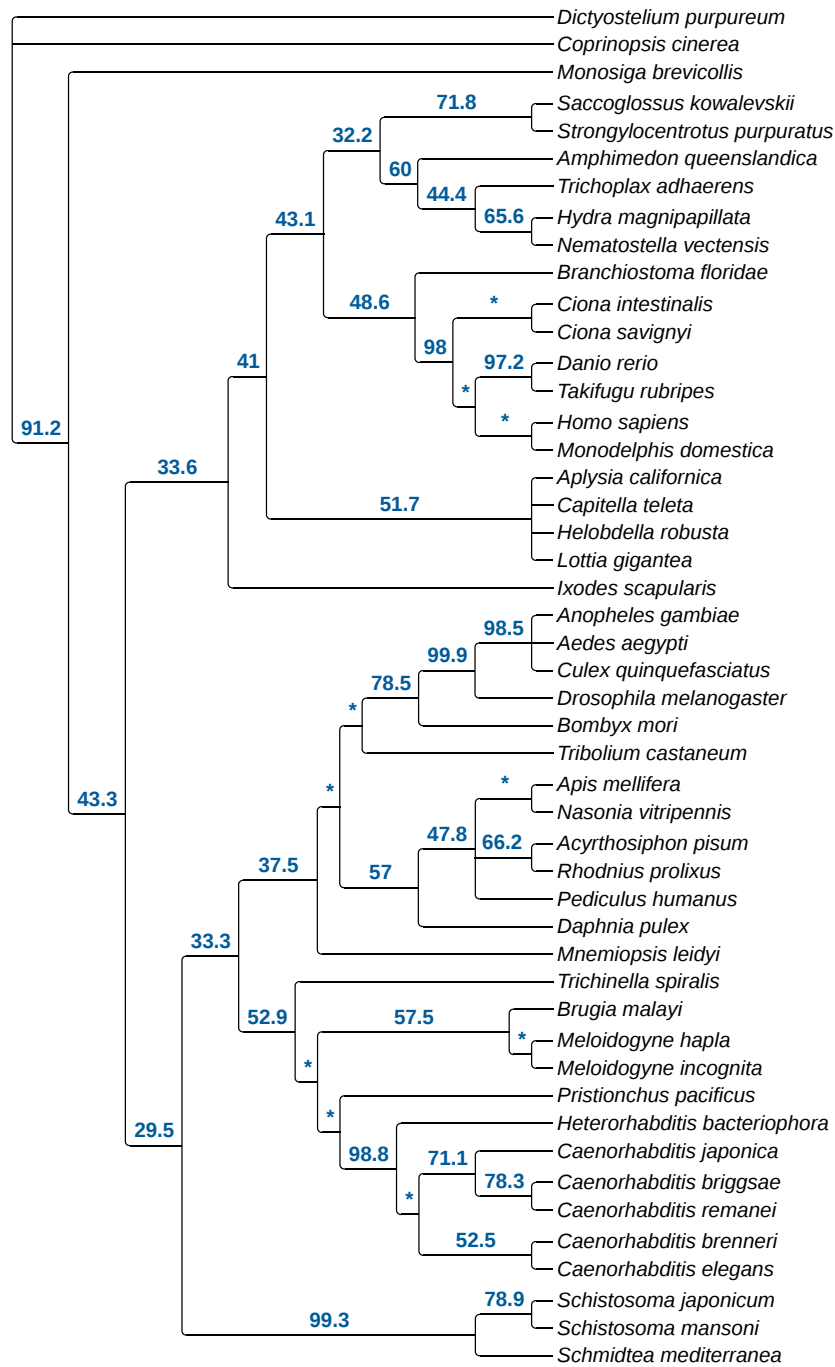
Figure S10: MP strict consensus tree as result of MP search with the subset of 7,630 non-overlapping parsimony-informative NIPs. Bootstrap support (in percent) is given above the corresponding branches (in blue). A star denotes 100% support.