Evolutionary origins of human apoptosis and genome stability gene networks

Supporting Online Material

Mauro A. A. Castro[1,3*], Rodrigo J. S. Dalmolin[1*], José C. F. Moreira[1], José C. M. Mombach[4] & Rita M. C. de Almeida[2]

[1]Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, Brazil. [2]Instituto de Física, Universidade Federal do Rio Grande do Sul, Avenida Bento Gonçalves 9500, Porto Alegre 91501-970, Caixa Postal 15051, Brazil. [3]Universidade Luterana do Brasil, Avenida Itacolomi 3600, Gravataí 94170-240, Brazil. [4]Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil.

[*]These authors contributed equally to this work.

Correspondence to: Mauro A. A. Castro[1] (mauro@ufrgs.br) or Rita M. C. de Almeida[2] (rita@if.ufrgs.br).

Contents

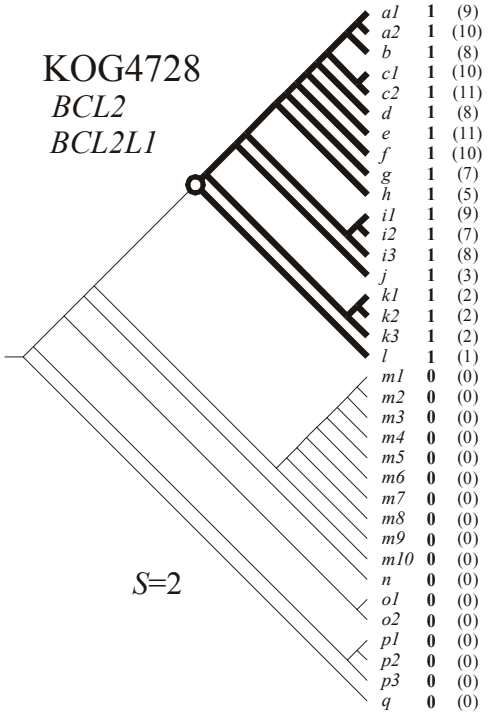
LIST OF SUPPLEMENTARY FIGURES

1. Supplementary results and discussion

1.1. Exemplification of the evolutionary analysis

1.1.1. Parsimony analysis and evolutionary scenarios

To illustrate the parsimony analysis, consider the evolutionary scenario presented in **Supplementary Figure S1** for KOG4728. This KOG comprises 123 genes distributed among 18 species and the pattern of presence or absence of the species in the analyzed set of species is indicated by, respectively, zero or one. Observe that at least one ortholog is present in all extant metazoan species from *Homo sapiens* (*a1*) to *Caenorhabditis elegans* (*l*).



Supplementary Figure S1. Inferring ancestral state of human apoptosis and genome stability genes. Parsimony analysis of KOG 4728. The inconsistency value *S* is indicated. Branch codes: ***a1*** (*Homo sapiens*); ***a2*** (*Pan troglodytes*); ***b*** (*Macaca mulatta*); ***c1*** (*Rattus norvegicus*); ***c2*** (*Mus musculus*); ***d*** (*Canis familiaris*); ***e1*** (*Bos Taurus*); ***f*** (*Monodelphis domestica*); ***g*** (*Gallus gallus*); ***h*** (*Xenopus tropicalis*); ***i1*** (*Takifugu rubripes*); ***i2*** (*Tetraodon nigroviridis*); ***i3*** (*Danio rerio*); ***j*** (*Ciona intestinalis*); ***k1*** (*Drosophila melanogaster*); ***k2*** (*Anopheles gambiae*); ***k3*** (*Apis mellifera*); ***l*** (*Caenorhabditis elegans*); ***m1*** (*Kluyveromyces lactis*); ***m2*** (*Saccharomyces cerevisiae*); ***m3*** (*Candida glabrata*); ***m4*** (*Eremothecium gossypii*); ***m5*** (*Debaryomyces hansenii*); ***m6*** (*Yarrowia lipolytica*); ***m 7*** (*Aspergillus fumigatus*); ***m 8*** (*Schizosaccharomyces pombe*); ***m9*** (*Filobasidiella neoformans*); ***m10*** (*Encephalitozoon cuniculi*); ***n*** (*Dictyostelium discoideum*); ***o1*** (*Arabidopsis thaliana*); ***o2*** (*Cyanidioschyzon merolae*); ***p1*** (*Plasmodium falciparum*); ***p2*** (*Cryptosporidium hominis*); ***p3*** (*Thalassiosira pseudonana CCMP1335*); ***q*** (*Giardia lamblia ATCC 50803*).

To explain this common genetic trace, the parsimonious evolutionary scenario assumes that the ortholog was present in the last common ancestor (LCA) of metazoans, and was genetically transmitted to the descendents up to the species we know today. Given that two human apoptotic genes are listed in KOG4728 (*BCL2* and *BCL2L1*), the evolutionary root of these two human genes is inferred in the origin of metazoans as well. The evolutionary scenarios for the remaining 141 KOGs are provided in **Supplementary Figures S14-S49**, and the parsimony analysis required to reconcile orthologous groups with the species tree topology is resolved according to the gain/penalty approach (Mirkin et al. 2003), where the most parsimonious scenario of presence/absence of all the genes at all ancestral nodes of the tree was predicted by

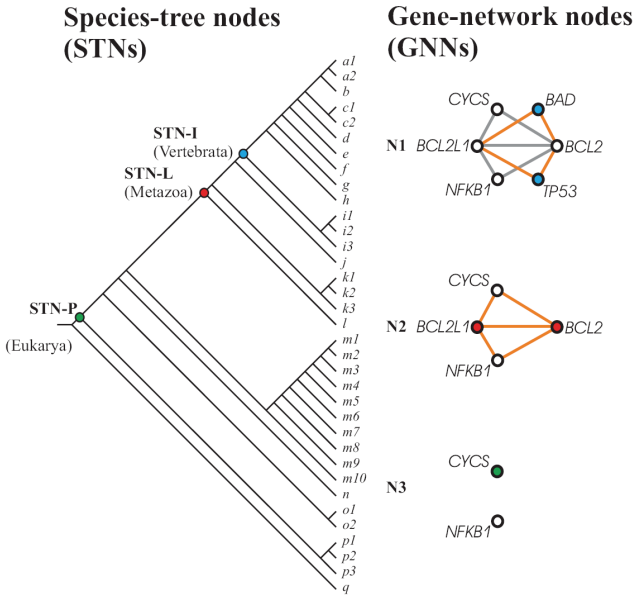
$$S = \lambda + g\gamma, \qquad (1)$$

where λ is the number of gene losses, γ is the number of gene gains and g is the gain penalty. Then, a parsimonious scenario must minimize the total score according to the lowest evolutionary cost. The minimal score is referred to as the inconsistency value S for the given orthologous group and is associated with the number/types of events required to reconcile the evolutionary scenario for the given orthologous group with the species tree. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (*i.e.* g=2), and one cost unit for gene loss (Snel et al. 2002; Kunin and Ouzounis 2003). Therefore, in the case of KOG4728 illustrated in **Supplementary Figure S1, S=2.**

1.1.2. From STNs to GNNs

In order to exemplify the analysis presented in Figure 3 in the paper, here we applied the same approach but in a small set of six apoptotic genes (**Supplementary Figure S2)**. Note that, as defined in the paper, any *species-tree node* (STN) represents a LCA, while any *gene-network node* (GNN) represents an ortholog in the human apoptosis and genome stability gene network (*i.e.* one STN can comprise more than one GNN roots). The orthology information from STNs is then overlaid on the network graphs, which are arranged in **Supplementary Figure S2** to optimally display the intersection within the species tree.

Supplementary Figure S2. Orthology information transferred for graphs. The orthology information from STNs is overlaid on GNNs, which are arranged for best observing the intersection within the species tree. It illustrates the projection of orthology information inferred for six orthologs. Roots of an ortholog: color nodes; presence of an ortholog: white nodes. Branch codes as in **Supplementary Figure S1**.



Therefore, the orthology information from STN-I is transferred for graph N1 and shows that *BAD* and *TP53* genes are rooted in the LCA of vertebrates (blue GNNs), while the other four genes are rooted prior to this species tree level (white GNNs). Also, in graphs N2 and N3 we show the orthology information inferred in STN-L and -P, respectively. Observe that *BAD* and *TP53* genes were removed from graph N2 because they are not placed in the LCA of metazoans. By the same reason, only *NFKB1* and *CYCS* genes are placed in graph N3, which have no links because these two ancient components are not functionally associated in the network-based model of human apoptosis and genome stability genes (Castro et al. 2007). Therefore, every STN intersection can produce an orthology projection onto the network graph. We remark that all orthology analysis should be restricted to the network components, given that the links represent association among human proteins. The strategy is then to follow the evolutionary steps that lead to the human gene/protein association network, *i.e.,* the roots of the human genes. In the paper we present the results for the set of 180 human genes.

1.1.3. Plasticity analysis

To exemplify the plasticity analysis, consider the KOG4728 illustrated in **Supplementary Figure S1**. As stated before, this KOG comprises 123 genes distributed in 18 species. However, the orthologous genes are not equally distributed, as long as an orthologous group may have more genes associated with some species than others. In theory, if a given orthologous group has only one ortholog in every organism of the species tree, then its diversity will be maximum (H_α=1.0). In contrast, if all genes are present in only one organism, then its diversity will be minimum (H_α=0.0); note that H_α zero will never be achieved, as long as at least three species are required to build an orthologous group (Tatusov et al. 2003). Thus, the former indicates a broad distribution in the species tree comparing to the latter, a possible measure of the evolutionary conservation. As a complementary quantity, we also estimate the abundance D_α, which is simply obtained by the ratio between the number of orthologous genes and the number of organisms. If an orthologous group has only one ortholog per organism, *i.e.*, D_α=1.0, then it indicates poor representation in any particular species. In contrast, KOGs with D_α>>>1.0 indicates many variants per organisms, a possible measure of the evolutionary plasticity. These statistics are also extended to Inparanoid database in order to assess a different data source.

1.2. Consistency of evolutionary and functional data

1.2.1. Network statistics

It is important to remark that some apoptotic genes are missing from KEGG map, which is used to construct the apoptosis gene network. However, given our network approach we can not describe genes individually. To consider other genes we must consider firstly the protein interaction network. This problem is critical for reliable protein network reconstruction and KEGG represent curated reconstructions of protein-protein interactions. This feature made KEGG a reference source that benchmarks numerous system biology databases (*e.g.* Cancer Genome Projected at NIH/CGAP and STRING database). Also, such option for KEGG

potentially enhances the primary applicability of our orthology map, that is, the transferability of functional information from several organisms to human. A complete description of the construction of the genome maintenance gene network is described in our previous work (Castro et al. 2007).

1.2.2. Orthologous groups statistics

Many of the current ortholog databases that uses genome-wide analysis will likely contain false-positives due to the limitations of the reciprocal-best-hits (RBH) approach (*e.g.* predict paralog as ortholog). However, in absence of golden standards, actually it is not possible to discriminate the best way of deriving orthologous groups, and whether there is in fact one way that works best for all applications. Although KOGs provide a sensible approach that does not rely on arbitrary score cut-offs, due to potential bias we extensively confronted our evolutionary scenarios obtained from KOGs with other independent sources (*i.e.* SGD, MGD, CGP, XP databases, as discusses in the paper, and Inparanoid database). Next, we present the evolutionary scenario for apoptosis and genome stability orthologs obtained from Inparanoid database, which provides a fully automatic method for finding orthologs (Remm et al. 2001).

1.2.2.1 Comparing evolutionary scenarios: KOG x Inparanoid orthologous groups

To test the robustness of the evolutionary scenarios predicted by our analysis we investigated the evolutionary roots of the same set of genes but using a different orthology detection approach. Here we consider the Inparanoid database. In contrast to KOG algorithm, Inparanoid is designed to find orthologs and in-paralogs between two species and to separate in-paralogs from out-paralogs (Remm et al. 2001).

In **Supplementary Figure S3A** we present the topology of the species tree for the available organisms at Inparanoid database. The evolutionary roots inferred for human apoptosis and genome stability genes are indicated in **Supplementary Figure S3B,C**. Comparing to the species tree derived from KOGs, Inparanoid derived species tree shows the same human ancestral nodes, except for the basal position. Also, the overall results of the pooled orthologs are very

similar comparing to the results presented in the paper (see Figure 2). To assess quantitatively the contrasts between KOG and Inparanoid evolutionary scenarios we also compared the inconsistency scores *S* (**Supplementary Figure S4**). For the entire gene set, *S*=4.39(±2.07; KOG analysis), and *S*=5.27(±2.48; Inparanoid analysis). However, this statistics estimates the inconsistence of the evolutionary scenarios considering the entire species tree. As long as the resulting species trees are slightly different between databases, then we should consider a complementary quantity in order to estimate the inconsistency of the roots inferred for each gene. The inconsistence of each gene root can be defined as

$$\Delta STN_i = \left| KOG_{STN} - Inparanoid_{STN} \right| \qquad (2)$$

where KOG_{STN} represents the species-tree node where is rooted a given gene *i*, according to KOG database, and $Inparanoid_{STN}$ represents the root of the same gene *i* according to Inparanoid database. The ΔSTN average value for the entire gene set *n* gives the evolutionary inconsistency score *R* of the evolutionary scenarios. *R* is given by

$$R = \frac{\sum_{i}^{n} \Delta STN_i}{n} \qquad (3)$$

In **Supplementary Figure S5** we present *R* for apoptosis and genome stability genes. This figure estimates the divergence between KOG and Inparanoid derived scenarios, that is, R=1.709 for apoptosis (approximately two STNs up- and down-ward to the rooting point in the species tree) and R= 0.807 for genome stability (approximately one STNs up- and down-ward).

To address qualitatively the differences among the databases, we reconstructed the network graphs of the evolutionary scenarios using the evolutionary roots inferred from Inparanoid database (**Supplementary Figure S6**). Accordingly, the two major increments in apoptosis network are aligned between KOGs and Inparanoid databases: the emergence of the apoptosis intrinsic pathway at the metazoan origin and the emergence of the extrinsic pathway at the vertebrata origin. Although not the same components are present in both scenarios, the overall results in the network topology are equivalent. As addressed in the paper, here the evolutionary scenario shows that the genome maintenance mechanisms could be divided into two major segments: i) the evolution of genome stability gene network placed in the basal position of the species tree and ii) the evolution of apoptosis gene network with components rooted throughout eukaryotic evolution. Likewise, the network core of both systems is rooted before the divergence

of metazoans, while GNNs placed in the periphery of the networks represent more recent evolutionary innovations. Further details for all Inparanoid orthologous groups, the evolutionary scenarios and the corresponding *S* value are presented in **Supplementary Figures S50** to **S94** and also provided in spreadsheet format (**Supplementary Table S4**).

1.2.3. Mouse and yeast statistics

Any evolutionary scenario providing the putative history that has given rise to the species placed in the phylogenetic tree must reflect, to some extent, the known functional differences and similarities observed in the living organism. Incongruence between evolutionary and functional data indicates a biased construction and should be revised. In the scenario described in the paper, the network regions of high and low evolutionary plasticity indicate that the network was not equally tolerant to genetic changes (*e.g.* mutations), as long as apoptosis have produced more variants of its components than genome stability. One possible explanation for the difference in plasticity between apoptosis and genome stability genes is to assume that the evolutionary plasticity of the network is proportional to the genetic changes that a living organism may support without disrupting its viability. Thus, to investigate this assumption, we assessed the lethality data of the unicellular eukaryotic model *Saccharomyces cerevisiae* available in the *Saccharomyces* Genome Database (SGD) (Hirschman et al. 2006), as well as the lethality data of *Mus musculus* orthologs in Mouse Genome Database (MGD) (Eppig et al. 2007).

Both databases provide gene knock-out data used to map genetic essentiality in the genomes. However, care should be taken when considering SGD together with MGD database. The available lethality statistics for this unicellular eukaryote comes from systematic experiments that cover almost all ORFs in the yeast genome, representing a comprehensive cellular lethality map. In contrast to yeast, mouse genes are not equally studied as long as *Mus musculus* data did not arise from systematic experiments. This explain why the network area that concentrates several yeast essential orthologs corresponds mainly to those in mouse that lacks knock-out data (see, in the paper, **Figure 5**: white GNNs). Also, the phenotypic statistics in MGD database consider lethal any allele that causes death anytime after fertilization and before the postnatal day 2; thus, knock-out alleles may indicate “developmental lethality” or “essentiality” to embryonic stem cells. Furthermore, evidences for mouse lethality data are obtained according to the

frequency expected by Mendelian genetics (*i.e.* zygosity and allelic distribution observed in the offspring): any significant deviation from the expected frequency for the knock-out allele indicates lethality, but it does not mean that the lethal genotype does not arise in the offspring.

1.2.4. Cancer statistics

In the same line discussed above, care should be taken in order to consider cancer statistics together with yeast and mouse due to differences among data sources. For instance, *CAN* gene statistics comes mainly from epidemiological data and shows exclusively genes in which mutations that are causally implicated in oncogenesis have been described at least in two independent reports, showing mutations in primary patient material (Futreal et al. 2004). According to CGP census, the underlying rationale for interpreting a mutated gene as causal in cancer development is that the number and pattern of mutations in the gene are likely to have been selected because they confer a growth advantage on the cell population from which the cancer has developed (Futreal et al. 2004). Also, in contrast to mouse and yeast knock-out alleles, *CAN* gene may have a range of mutations, from a single nucleotide substitution to a complete transcript disruption (*i.e.* null alleles).

In order to circumvent such data limitations and improve the analysis we further investigated the human statistics assessing the genotypic profile of several *CAN* gene loci. We attempt to obtain the proportion of null and non-null alleles in human following the strategy used in mouse to infer lethality according to the expected frequency in a Mendelian distribution. We focus the analysis in the set of *CAN* genes placed in ρ module, given that they are collectively represented in the same locus-specific mutation database – XP mutation database (http://www.xpmutations.org). These *CAN* genes are also associated with the same DNA repair function (nucleotide-excision repair) and are related to three rare autosomal recessive human clinical disorders (Xeroderma pigmentosum, Cockayne Syndrome and Trichothiodystrophy), which may turn reliable the obtaining of a representative human sample. Due to obvious reason, human data do not come from uniform samples but represent a historical collection (*e.g.* case reports documented all around the world). We retrieved 182 mutated genotypes available in that database, which is then pooled according to the zygosity and the presence of null and non-null alleles (see Table 1A in the paper). Sample number is also compared to a second database in

order to attest the representativeness of the database (see **Table 1B** in the paper). In **Supplementary Figure S7** we present the cancer statistics according to the presence of somatic and/or germline *CAN* genes in the network (*i.e.* apoptosis and genome stability). In order to inspect these results together with the main results present in the paper, this figure also shows the statistical contrasts for plasticity data in relation to abundance $D\alpha$, diversity $H\alpha$, gene function, and gene essentiality.

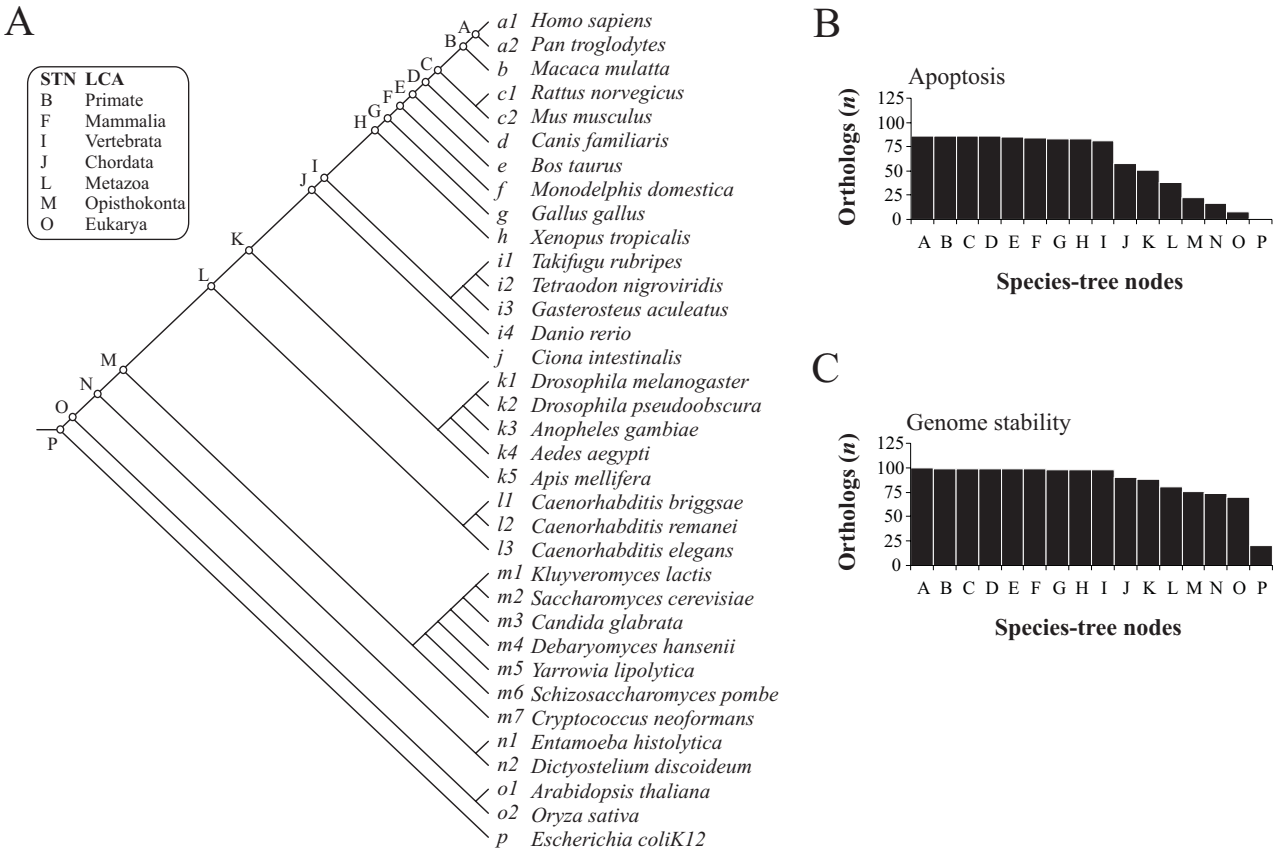
1.3. The deep root of eukaryotes

The deep root of eukaryotes is controversial, showing low resolution, and thus should be considered with caution (Doolittle 1999; Baldauf 2003). For instance, *Giardia lamblia* is an extant species that diverged in the consensus tree at STN-Q (eukarya). Due to this early divergence, *Giardia lamblia* carries important information on the evolutionary tree. However, Giardia is a parasite, prone to gene loss (Best et al. 2004). This means that its ancestors may have had more genes that their parasite descendents. In this case information is lost on the number of orthologs inferred in the basal position of the species tree. To bypass this limitation it has been suggested that *Giardia lamblia* and other related organisms should be jointly considered to constitute the representative lineage of the descendent species: the gene loss would mutually compensate if genes were lost differentially by different species (Makarova et al. 2005). Accordingly, in a combined scenario (*e.g.* STN-P and -Q), the contrast describe in the paper between apoptosis and genome stability pathways could be extended to the origins of all eukaryotes in the species tree. Additional evidence can be inferred considering the likely origin of the ancestral eukaryotic KOGs by identifying their closest prokaryotic orthologous groups (COGs). The KOG-to-COG correspondence is presented in **Supplementary Figure S8** and **Supplementary Table S2**, and shows that 77.0 % of the genome stability orthologs have identifiable prokaryotic orthologous groups, against 39.5% for apoptotic orthologous genes. Furthermore, this scenario is consistent with the origins described for a small set of COGs that can be traced back to the universal ancestor and that recapitulate the three-domain phylogeny (Harris et al. 2003), since in our list only eukaryotic orthologous groups associated with genome

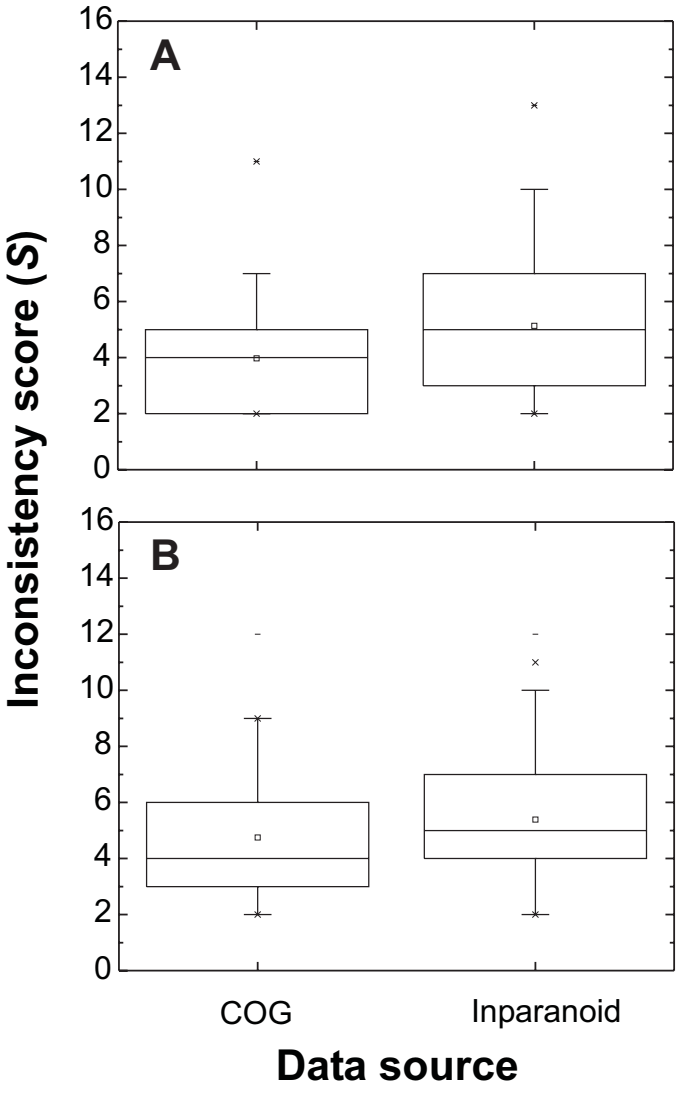
stability functions match these universally conserved COGs (**Supplementary Figure S8**, red stripes).

1.4. The deep root of metazoans

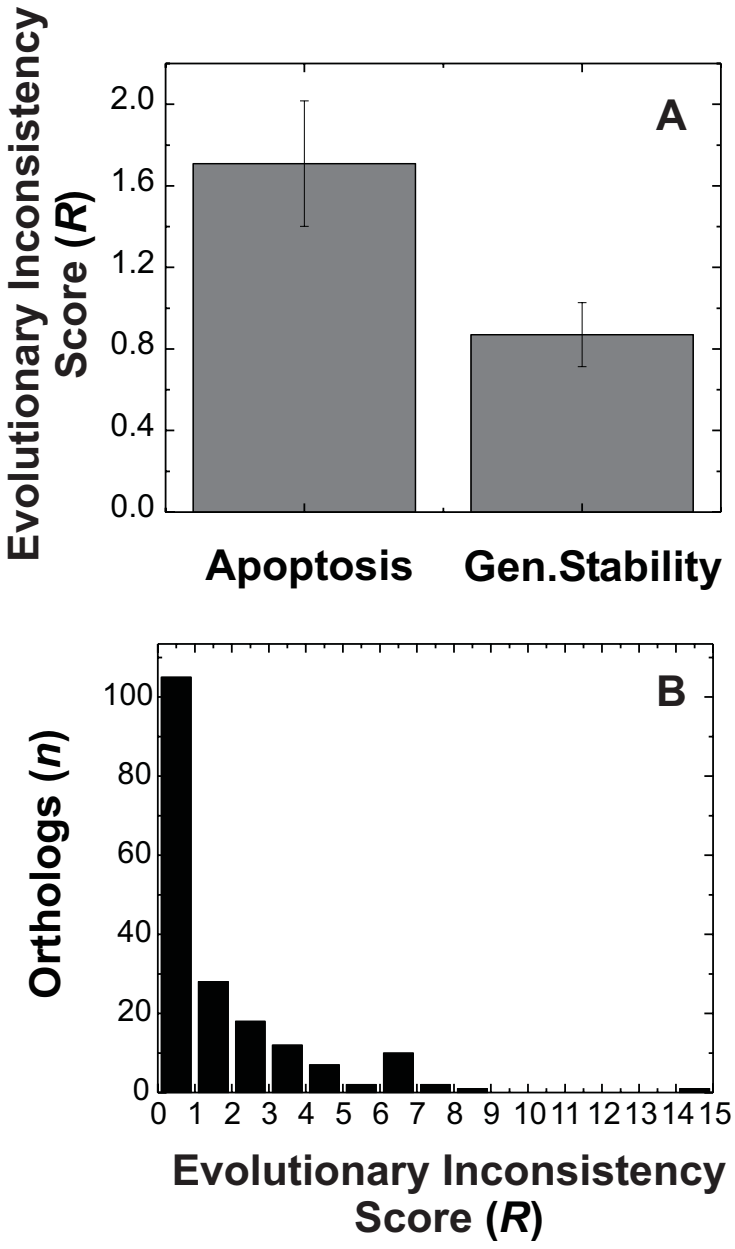
The deep root of metazoan is another controversial issue. However, the current work do not propose the phylogeny of the organisms; instead, here is considered an integration of a variety of phylogenies (Katinka et al. 2001; Pennisi 2003; Baldauf 2003; Delsuc et al. 2005; Ciccarelli et al. 2006; Letunic and Bork 2007). As mentioned in the *Material and Methods*, for each orthologous group associated with the human apoptosis and genome stability genes, our problem is how to find the earliest ortholog in the eukaryote phylogeny (*i.e.* given the species tree, where the human gene roots is placed). Therefore, we focused in the interpretation of our data given the already described species tree topology. Despite these limitations, a different phylogeny in which *C.elegans* isn't at the metazoan root would be welcome, since this species tree node showed the origin of many orthologs of the human apoptosis gene network, which could potentially affect the interpretation of our results. Hence, we carefully included *Nematostella vectensis* as a subsample, which thus changes the base of metazoa (**Supplementary Figure S9**). Comparing this figure with the original analysis (**Figure 2B**), the distribution of orthologs remains almost the same (the complete analysis is available at **Supplementary Table S5**). The presence or absence in *Nematostella* was unknown for 34 KOGs, in which case a change from *C. elegans* was assumed (and in each case this meant assuming the presence of the KOG absent in *C. elegans*). The result after this process is that only 1 KOG was found to be different between *C. elegans* and *Nematostella*, however, even assuming all 34 KOGs where *Nematostella* data were unavailable were also different from *C. elegans* did not change the overall pattern (see Supplementary Figure S9).



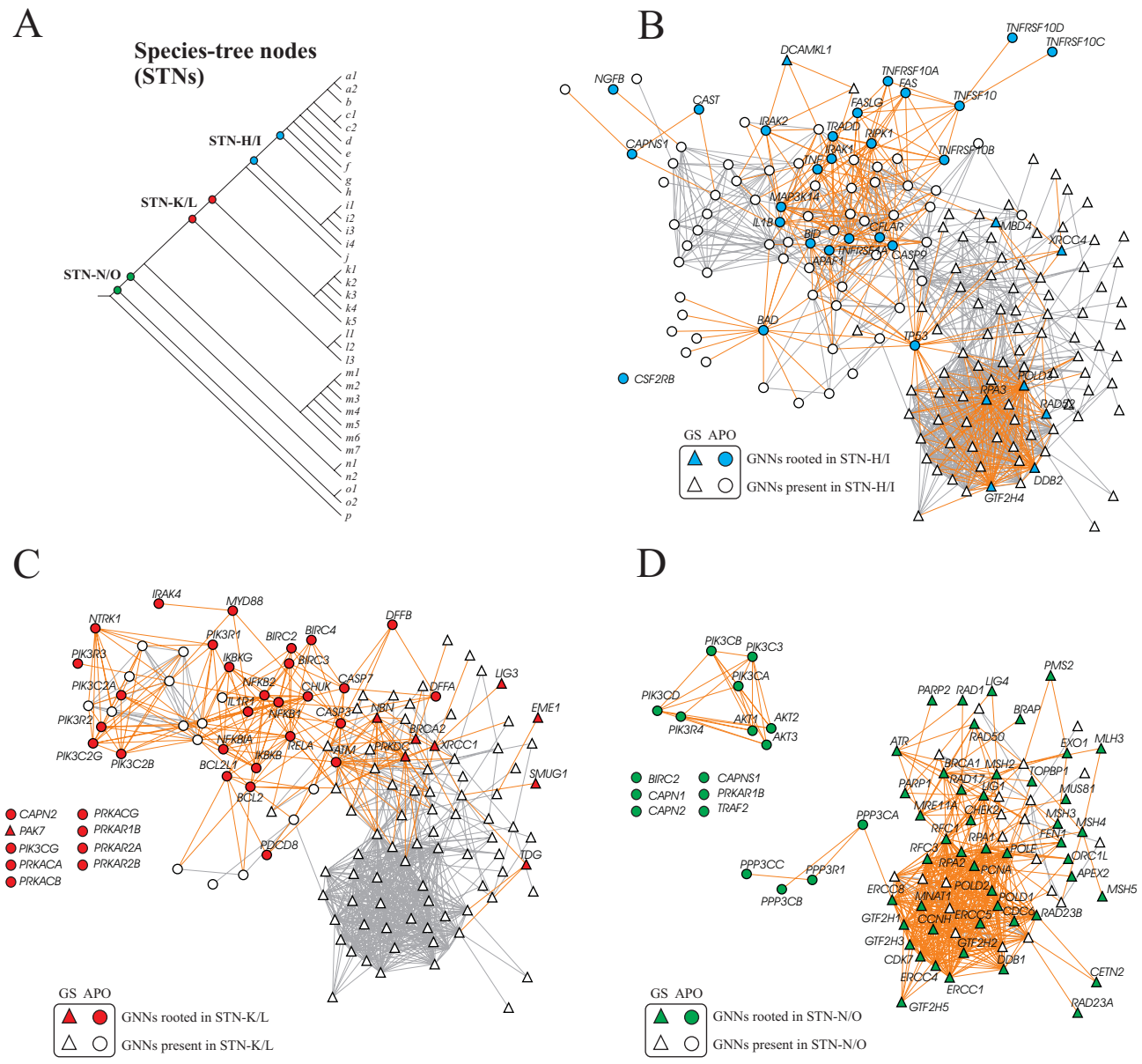
Supplementary Figure S3. Inferring ancestral state of human apoptosis and genome stability genes according to Inparanoid Database. (**A**) Eukaryote species tree topology used in the parsimony analysis. Species-Tree Nodes (STNs) and the corresponding Last Common Ancestor (LCA) are indicated. (**B**) Distribution of apoptosis orthologs according to the roots inferred in the species tree. (**C**) Distribution of genome stability orthologs, as in B.



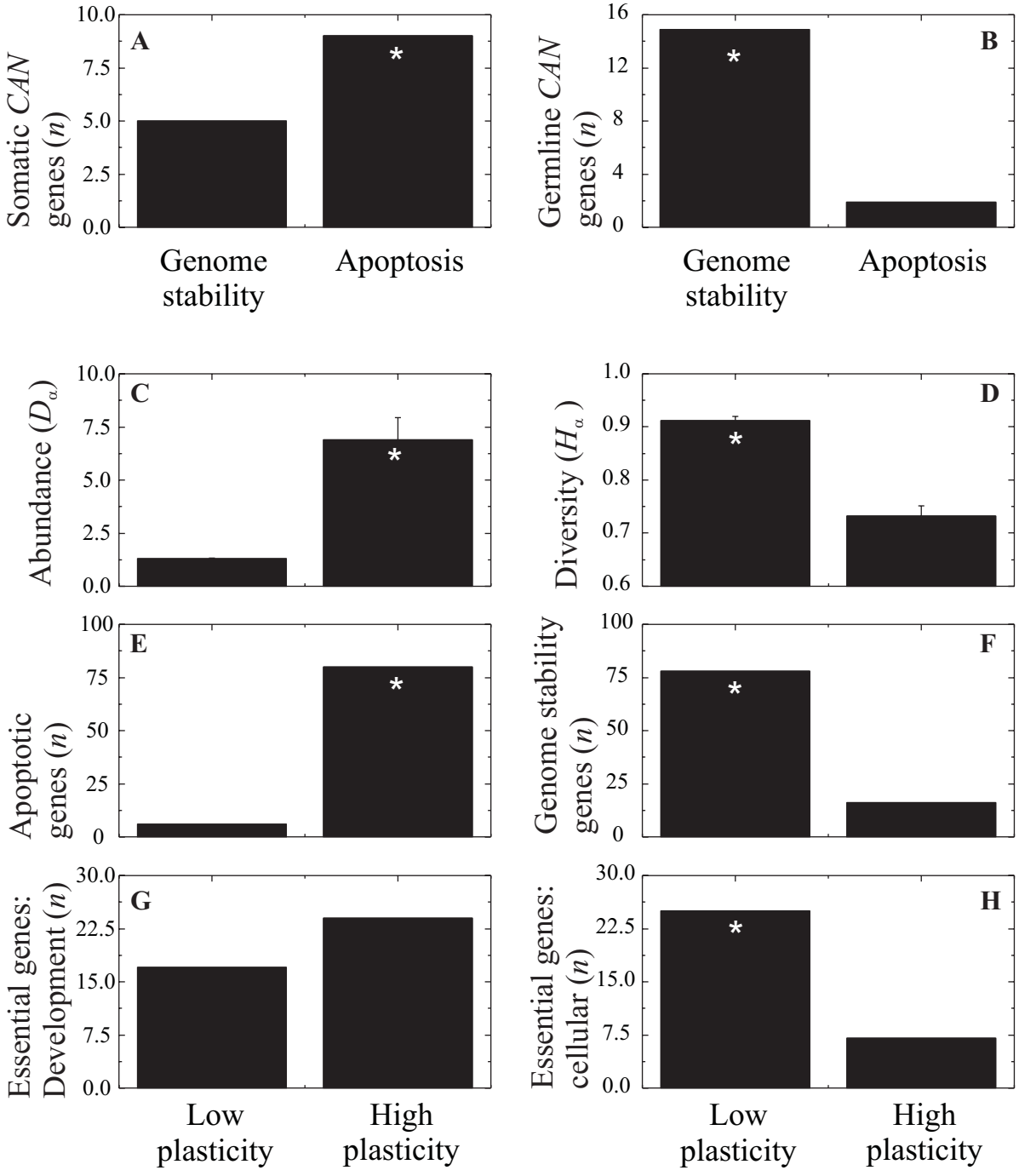
Supplementary Figure S4. Inconsistency scores *S* estimated from KOG and Inparanoid orthology analysis. Apoptosis (**A**) and genome stability gene set (**B**). For the entire gene set, S=4.39(±2.07; KOG analysis), and S=5.27(±2.48; Inparanoid analysis).



Supplementary Figure S5. Comparing the evolutionary roots inferred from COG and Inparanoid orthology analysis. (**A**) Apoptosis and genome stability gene sets: evolutionary inconsistency score *R* is presented as mean ±SEM. (**B**) Frequency distribution of orthologs for the entire gene set according to *R*.



Supplementary Figure S6. Inparanoid evolutionary scenarios: from species-tree nodes (STNs) to gene-network nodes (GNNs). The orthology information from STNs (**A**) is overlaid on GNNs (**B, C** and **D**), which are arranged to optimally display the intersection within the species tree. Roots of an ortholog: color nodes; presence of an ortholog: white nodes. B: Projection of STN-H/I; C: Projection of STN-K/L; D: Projection of STN-O/N.



Supplementary Figure S7. Summary of cancer statistics. Contrast between gene functions according to the presence of somatic **(A)** and germline **(B)** *CAN* genes. In order to observe the main results present in the paper, we also show the contrasts between plasticity data according to abundance *D*α **(C)**, diversity *H*α **(D)**, gene function **(E, F)**, and gene essentiality **(G, H)**. Network plasticity groups follow the categories defined in **Fig.4C**: Low plasticity groups contain those genes described in class-a gene set, while high plasticity groups contain those genes described in class-a/c gene set.*P<0.05 (Mann-Whitney U Test).

Apoptosis

Gene Symbol	KOG-toCOG
BAD	NOG10429
BAX	NOG16176
BCL2	KOG4728
BCL2L1	KOG4728
BID	NOG10549
BIRC2	KOG1101
BIRC3	KOG1101
BIRC4	KOG1101
CAPN1	KOG0045
CAPN2	KOG0045
CAPNS1	KOG0037
CASP10	KOG3573
CASP3	KOG3573
CASP6	KOG3573
CASP7	KOG3573
CASP8	KOG3573
CASP9	KOG3573
CAST	KOG1181
CFLAR	KOG3573
CSF2RB	NOG04828
DFFA	NOG04137
DFFB	NOG05130
FADD	NOG10546
FAS	NOG10520
FASLG	NOG07555
IKBKG	KOG0161

Gene Symbol	KOG-toCOG
IL1A	NOG08276
IL1B	NOG09287
IL1R1	NOG04905
IL1RAP	KOG4641
IL3	NOG21098
IL3RA	NOG21096
MYD88	NOG07319
NGFB	NOG07040
PIK3R1	KOG4637
PIK3R2	KOG4637
PIK3R3	KOG4637
PIK3R4	KOG1240
PIK3R5	NOG05250
PPP3R1	NOG04634
PRKAR1B	KOG1113
RELA	NOG04893
TNF	NOG08240
TNFRSF10A	NOG13097
TNFRSF10B	NOG13097
TNFRSF10C	-
TNFRSF10D	NOG36564
TNFRSF1A	NOG06963
TNFSF10	NOG08316
TP53	NOG07483
TRADD	NOG04168
TRAF2	KOG0297

Gene Symbol	KOG-toCOG
AKT1	KOG0690 / COG0515
AKT2	KOG0690 / COG0515
AKT3	KOG0690 / COG0515
APAF1	KOG4155 / COG2319
ATM	KOG0892 / COG5032
CHUK	KOG4250 / COG0515
CYCS	KOG3453 / COG3474
IKBKB	KOG4250 / COG0515
IRAK1	KOG1187 / COG0515
IRAK2	KOG1187 / COG0515
IRAK4	KOG1187 / COG0515
MAP3K14	KOG0198 / COG0515
NFKB1	KOG0504 / COG0666
NFKB2	KOG0504 / COG0666
NFKBIA	KOG0504 / COG0666
NTRK1	KOG1026 / COG0515
PDCD8	KOG1346 / COG0446
PIK3C2G	KOG0905 / COG5032
PIK3CA	KOG0904 / COG5032
PPP3CA	KOG0375 / COG0639
PPP3CB	KOG0375 / COG0639
PPP3CC	KOG0375 / COG0639
PRKACA	KOG0616 / COG0515
PRKACB	KOG0616 / COG0515
PRKACG	KOG0616 / COG0515
PRKAR2A	KOG1113 / COG0664

Gene Symbol	KOG-toCOG
PRKAR2B	KOG1113 / COG0664
RIPK1	KOG0192 / COG0515
PIK3C2A	KOG0905 / COG5032
PIK3C2B	KOG0905 / COG5032
PIK3C3	KOG0906 / COG5032
PIK3CB	KOG0904 / COG5032
PIK3CD	KOG0904 / COG5032
PIK3CG	KOG0904 / COG5032

KOGs that have a counterpart in prokaryotic groups

KOGs without a counterpart in prokaryotic groups

KOGs that match universally conserved COGs

Genome stability

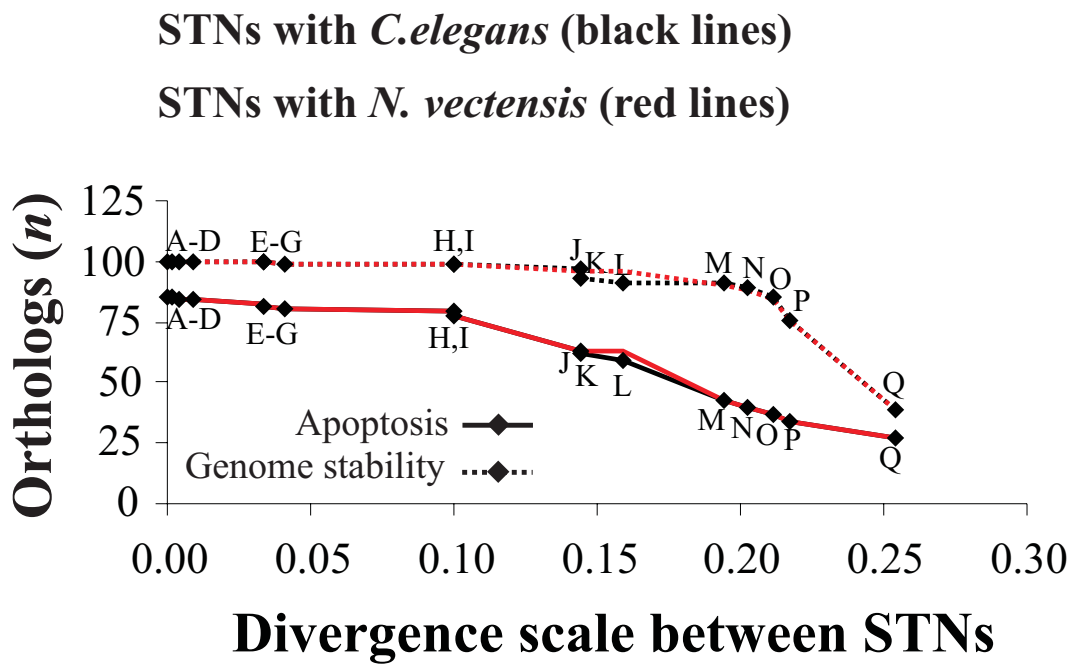
Gene Symbol	KOG-toCOG
BRAP	KOG0804
BRCA1	KOG4362
BRCA2	KOG4751
DDB1	KOG1897
EME1	NOG05923
GTF2H1	KOG2074
GTF2H5	KOG3451
MBD4	KOG4161
NBN	NOG06900
PARP1	KOG1037
PARP2	KOG1037
POLD3	NOG05166
POLD4	NOG12441
RAD1	KOG3194
RPA3	NOG10964
SHFM1	KOG4764
SMUG1	NOG04402
TOPBP1	KOG1929
XAB2	KOG2047
XRCC1	KOG3226
XRCC4	NOG07367
XRCC5	KOG2326
XRCC6	KOG2327

Gene Symbol	KOG-toCOG
APEX1	KOG1294 / COG0708
APEX2	KOG1294 / COG0708
ATR	KOG0890 / COG5032
CCNH	KOG2496 / COG5333
CDC6	KOG2227 / COG1474
CDK7	KOG0659 / COG0515
CETN2	KOG0028 / COG5126
CHEK1	KOG0590 / COG0515
CHEK2	KOG0615 / COG0515
DCAMKL1	KOG3757 / COG0515
DCLRE1C	KOG1361 / COG1236
DDB2	KOG4328 / COG2319
DMC1	*KOG1434 / COG0468
ERCC1	KOG2841 / COG5241
ERCC2	KOG1131 / COG1199
ERCC3	KOG1123 / COG1061
ERCC4	KOG0442 / COG1948
ERCC5	*KOG2520 / COG0258
ERCC6	KOG0387 / COG0553
ERCC8	KOG4283 / COG2319
EXO1	*KOG2518 / COG0258
FEN1	*KOG2519 / COG0258
GTF2H2	KOG2807 / COG5151
GTF2H3	KOG2487 / COG5242
GTF2H4	KOG3471 / COG5144

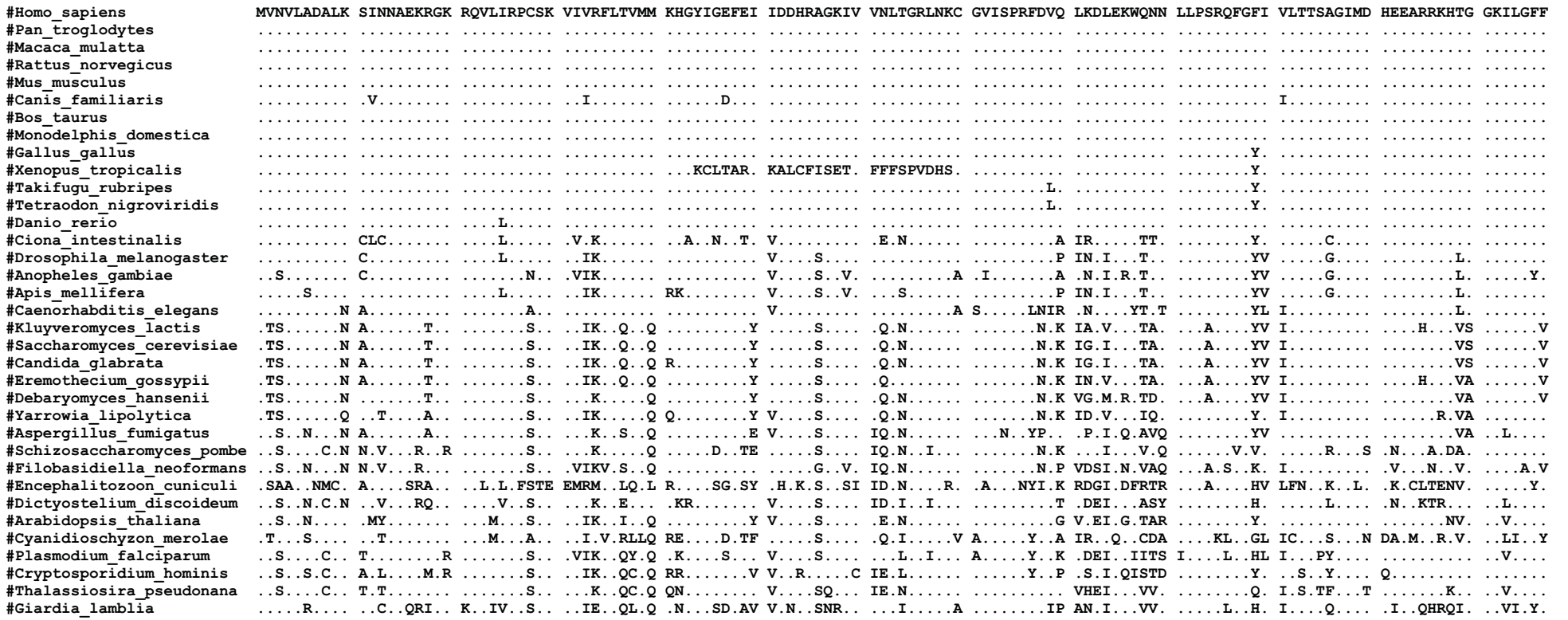
Gene Symbol	KOG-toCOG
LIG1	KOG0967 / COG1793
LIG3	KOG4437 / COG1793
LIG4	KOG0966 / COG1793
MLH1	KOG1979 / COG0323
MLH3	KOG1977 / COG0323
MNAT1	KOG3800 / COG5220
MPG	KOG4486 / COG2094
MRE11A	KOG2310 / COG0420
MSH2	KOG0219 / COG0249
MSH3	KOG0218 / COG0249
MSH4	KOG0220 / COG0249
MSH5	KOG0221 / COG0249
MSH6	KOG0217 / COG0249
MUS81	KOG2379 / COG1948
MUTYH	KOG2457 / COG1194
NTHL1	KOG1921 / COG0177
OGG1	KOG2875 / COG0122
ORC1L	KOG1514 / COG1474
PAK7	KOG0578 / COG0515
PCNA	*KOG1636 / COG0592
PMS2	KOG1978 / COG0323
PMS2L3	KOG1978 / COG0323
PNKP	KOG2134 / COG0241
POLD1	KOG0969 / COG0417
POLD2	KOG2732 / COG1311

Gene Symbol	KOG-toCOG
PRKDC	KOG0891 / COG5032
RAD17	KOG1970 / COG0470
RAD23A	KOG0011 / COG5272
RAD23B	KOG0011 / COG5272
RAD50	KOG0962 / COG0419
RAD51	*KOG1434 / COG0468
RAD52	KOG4141 / COG5055
RAD54B	KOG0390 / COG0553
RAD54L	KOG0390 / COG0553
RFC1	KOG1968 / COG0470
RFC3	KOG2035 / COG0470
RFC4	KOG0989 / COG0470
RFC5	KOG0990 / COG0470
RPA1	KOG0851 / COG1599
RPA2	KOG3108 / COG5235
TDG	KOG4120 / COG3663
UNG	KOG2994 / COG0692
XPA	KOG4017 / COG5145
XPC	KOG2179 / COG5535
PIK3C2A	KOG0905 / COG5032
PIK3C2B	KOG0905 / COG5032
PIK3C3	KOG0906 / COG5032
PIK3CB	KOG0904 / COG5032
PIK3CD	KOG0904 / COG5032
PIK3CG	KOG0904 / COG5032

Supplementary Figure S8. KOG-to-COG correspondence. Orthology information of the Clusters of Orthologous Groups (COGs) was retrieved through the KOG-to-COG assignments in the STRING database. Red stripes highlight the eukaryotic groups that match the universally conserved prokaryotic groups in the KOG-to-COG correspondence according to Harris *et al.* (2003).



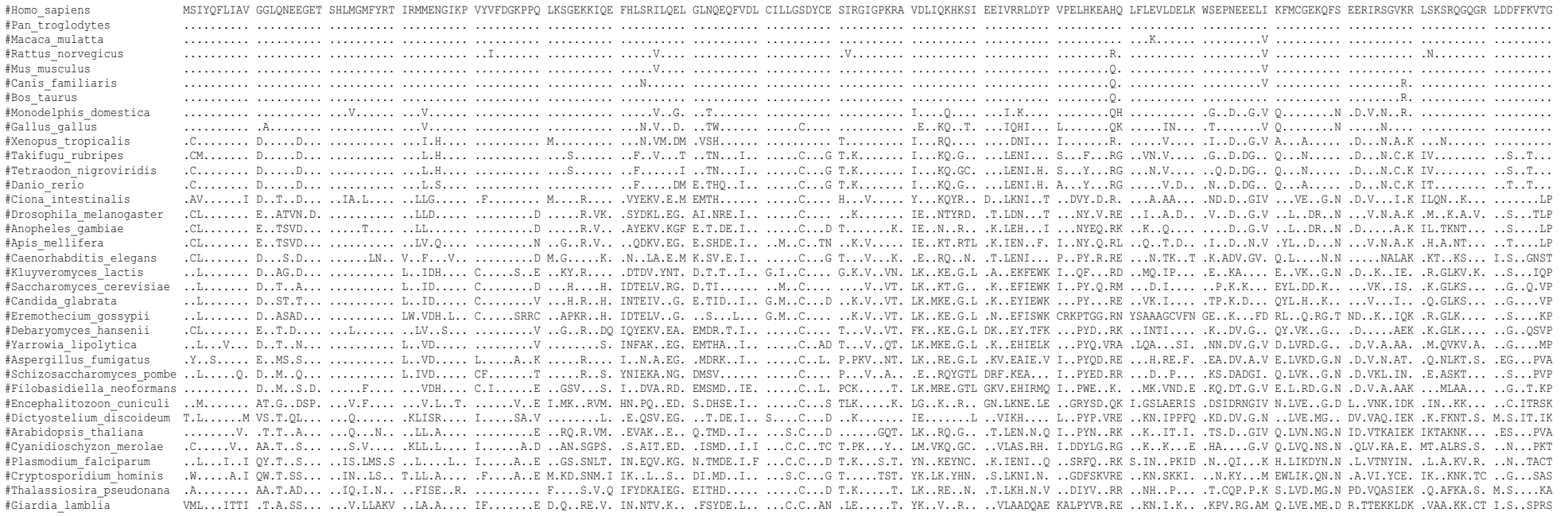
Supplementary Figure S9. Subsample of the data where *C. elegans* isn't at the root of metazoa. This figure presents a comparison between the original analysis (see **Figure 2**) and an alternative construction where *Nematostella vectensis* is manually placed in the position of *C.elegans*, according to the same species-tree nodes (STNs). Orthology data derive from *Nematostella vectensis* genome web site (http://genome.jgi-psf.org/Nemve1/Nemve1.home.html). (Data available at **Supplementary Table S5**).



Supplementary Figure S10. Comparison of amino acid sequences of 35 eukaryotic 40S ribosomal proteins obtained from a three-domain orthologous groups (COG0096/KOG1754). Amino acid sequences were aligned using ClustalW. All sites containing alignment gaps and missing-information were removed before computing the distance matrix. Dots represent amino acid residues that are identical to the corresponding sites of the first sequence (*i.e. H.sapiens* sequence). Protein IDs: *H.sapiens* (ENSP00000318646); *P.troglodytes* (ENSPTRP00000013350); *M.mulatta* (ENSMMUP00000007395); *R.norvegicus* (ENSRNOP00000024678); *M.musculus* (ENSMUSP00000008827); *C.familiaris* (ENSCAFP00000016520); *B.taurus* (ENSBTAP00000027627); *M.domestica* (ENSMODP00000007323); *G.gallus* (ENSGALP00000010942); *X.tropicalis* (ENSXETP00000006430); *T.rubripes* (NEWSINFRUP00000151157); *T.nigroviridis* (GSTENP00022669001); *D.rerio* (ENSDARP00000007879); *C.intestinalis* (ENSCINP00000008963); *D.melanogaster* (CG2033-PE); *A.gambiae* (ENSANGP00000029176); *A.mellifera* (ENSAPMP00000009198); *C.elegans* (F53A3.3.3); *K.lactis* (KLLA0B07601g); *S.cerevisiae* (YJL190C); *C.glabrata* (CAGL0K04587g); *E.gossypii* (AEL151C); *D.hansenii* (DEHA0B06864g); *Y.lipolytica* (YALI0D05731g); *A.fumigatus* (Afu1g15730); *S.pombe* (SPAC22A12.04c); *F.neoformans* (CNK02900); *E.cuniculi* (ECU09_1350); *D.discoideum* (DDB0167031); *A.thaliana* (AT5G59850.1); *C.merolae* (CMI202C); *P.falciparum* (MAL3P6.30); *C.hominis* (Chro.80500); *T.pseudonana* (26367); *G.lamblia* (15228).

```
#Homo_sapiens              MSTSKTGKHG HAKVHLVGID IFTGKKYEDI CPSTHNMDVP NIKRNDYQLI CIQDYLSLLT ETGEVRDLKL PEGELGKEIE GKYNAEDVQV SVMCAMSEEY AVAIK
#Pan_troglodytes           .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Macaca_mulatta            .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Rattus_norvegicus         .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Mus_musculus              .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Canis_familiaris          .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Bos_taurus                .......... .......... .......... .......... ......F... G........Q DS......R. ...D...... Q..DC.EILI T.LS..T..A .....
#Monodelphis_domestica     .......... .......... .......... .......... .......... .......... .......... .......... .......... .......... .....
#Gallus_gallus             .......... .......... ..N....... .......... .......... G......... .S........ ...D..... ..F......I ..IS..N..C .....
#Xenopus_tropicalis        .......... .......... .......... .......... ....C..... G.N......S DS.D.....M ...D..R..L M.HDG.EIL. T.LT..N..C ...L.
#Takifugu_rubripes         .......... .......... ..S....... .......... ....M..... G........D DN.......I .D.D...... N.NE..EILI T.LN..N..A .ISV.
#Tetraodon_nigroviridis    .......... .......... ..S....... .......... ....M..... G........D DN.......I .D.D...... S.HE..EILI T.LN..N..A .ISV.
#Danio_rerio               .......... .......... ...N...... .......... .........V D.SEF...MM DN.D....RV .D.D...... N.FA..EML. T.LS..G..S ...L.
#Ciona_intestinalis        .......... .....M.... ..DS...... S......Q.. ....TEV.VV DVA.QVTIME .D.DT.E..V E.KPMLDS.H KIIESG.CYI TILA.LGR.K VITA.
#Drosophila_melanogaster   .......... .....M.... ..SN...... .......... .V..E.L... A.D.F.T.M. .S.DL....V ......EQLR LDFDSK.LLC T.LK.CG..C VI...
#Anopheles_gambiae         .......... .....M.... .......... .......... .V..E....T D.D.F.V.MS DN.DL....I .D....TQLR SEFDSKEIVC T.LKSCG..T VI...
#Apis_mellifera            .......... .......... ...S...... .......... FV..E....A D.D...C.MA DN..L....I .D....AQLR ADHE.KELLC T.LK.CG..V VI...
#Caenorhabditis_elegans    .......... .....M.A.. ...S..L... .......... VV..RE.L.M A.D..C..MD PEC.QK.... .DT...QQ.R DA.EK.S.L. Q.VS.IG..A ILGW.
#Kluyveromyces_lactis      .......... .......A.. ...N..L..L S......E.. VV..TEF..L D.D.F...M. MD..TK.VRA ......DN.Q AAFDEG.LM. TIIA..G..A .ISF.
#Saccharomyces_cerevisiae  .......... .......TL. ......L..L S.....LE.. FV..SE...L D.D.....M. MD..TK.V.A ......DSMQ AAFDEG.LM. TIIS..G..A .ISF.
#Candida_glabrata          .......... .......A.. ......L..L S......E.. VV..TEF..L D.D.F...MN MD.DTK.VRH ......DQMQ AAFDEG.LM. TIIA..G..A .ISF.
#Eremothecium_gossypii     .......... .......A.. ......L..L S......E.. VV..TE...L D.D.F...M. MD..TK.VRA .D....ETMQ ASFDEG.LM. T.ITS.G..A .ISF.
#Debaryomyces_hansenii     .......... .......A.. ......L..L S......E.. .VG.REF..L D.D.F...M. ND.DTK.V.V ......DK.Q SEFDEG.LL. TIVS..G..A .ISF.
#Yarrowia_lipolytica       .......... .......AT. ...N..L..L S......E.. .V..EEF... D.D.F...F. AD.STK.VP. ....I.DK.Q TEFDEG.LI. T.IS..G..A VISY.
#Aspergillus_fumigatus     .......... .....I.AL. ......L..L S......... .VT.RE...L DVD.F...MD .A.NTK.V.. ....V.ER.Q KMF.EG.CN. TILT..G.QA CMDV.
#Schizosaccharomyces_pombe .......... .....I.AL. ..N.R....M S......... VV..DE...V N.D...N.M. TD.TTK.VR. ......N... EGFEEG.LII T.VS..G..I .L.CR
#Filobasidiella_neoformans .......... .......A.. ......L... S......... .VR.QEF..L D...F.N.MD SD.GSK.V.V ..T.I.QQ.M ADFE.G.LM. TIIS..D..Q .ISY.
#Encephalitozoon_cuniculi  T.SV.N.... A..TTISSKI LS..SNHKG. YTANDSII.C RPEKVQLK.. D.TSTFTDSS GSDS.DAGRM SSEDKIVQTV EGS.SS.LSL R.LPDFYKLE S.RPS
#Dictyostelium_discoideum  .......... ....NITA.. ........E. ......I... .VS.KE.TVM DV.......D AG...K..A. ..DDI....T QMLKEKEPL. ..IS.LGK.G V.SV.
#Arabidopsis_thaliana      V......... ...C.F.A.. ...S..L... V..S..C... HVN.T..... D.E..V.... DN.STK.... .NDD.LQQ.K SGFDDK.LV. ...S..G..Q IN.L.
#Cyanidioschyzon_merolae   V......... ...ANI..L. .........V A.TS.T.YQ. VVV.SE.... D.AEFC..MD DK..T..... DPDNVHAK.K EDFE.KQLT. V.LK..KQ.K IMQS.
#Plasmodium_falciparum     Y......... ...A.I.... ....R..... ..TS...... VV..TEL... D.E.FV...Y DN.DTK..S. .KDTVA.Q.R NLFDNKS.L. ..LS.CGQ.K II.A.
#Cryptosporidium_hominis   .......... ...A.I..L. ...N.....V ..TS...A.. FV..SE.... D.DNFT.... .N.STK..A. .TDAVALQ.Q QM.ADKAIL. G.LA.CGI.K I.SA.
#Thalassiosira_pseudonana  I.V....... ...CNFTAV. ......L..M V..S.GTT.. IVT.TEWEI. D.EEE.T.MD .G.NQK.VN. .GVPMAQ..R DAW.D.Q.S. T.QA.VGQ.Q V..Y.
#Giardia_lamblia           I......... ...CSITAV. .........S ...S..IMC. .VS.T..L.. N.D..AFYFD QNA.QQR.P. ENE..VADLR AAFE.DTIMI NIQR..NT.G I..F.
```

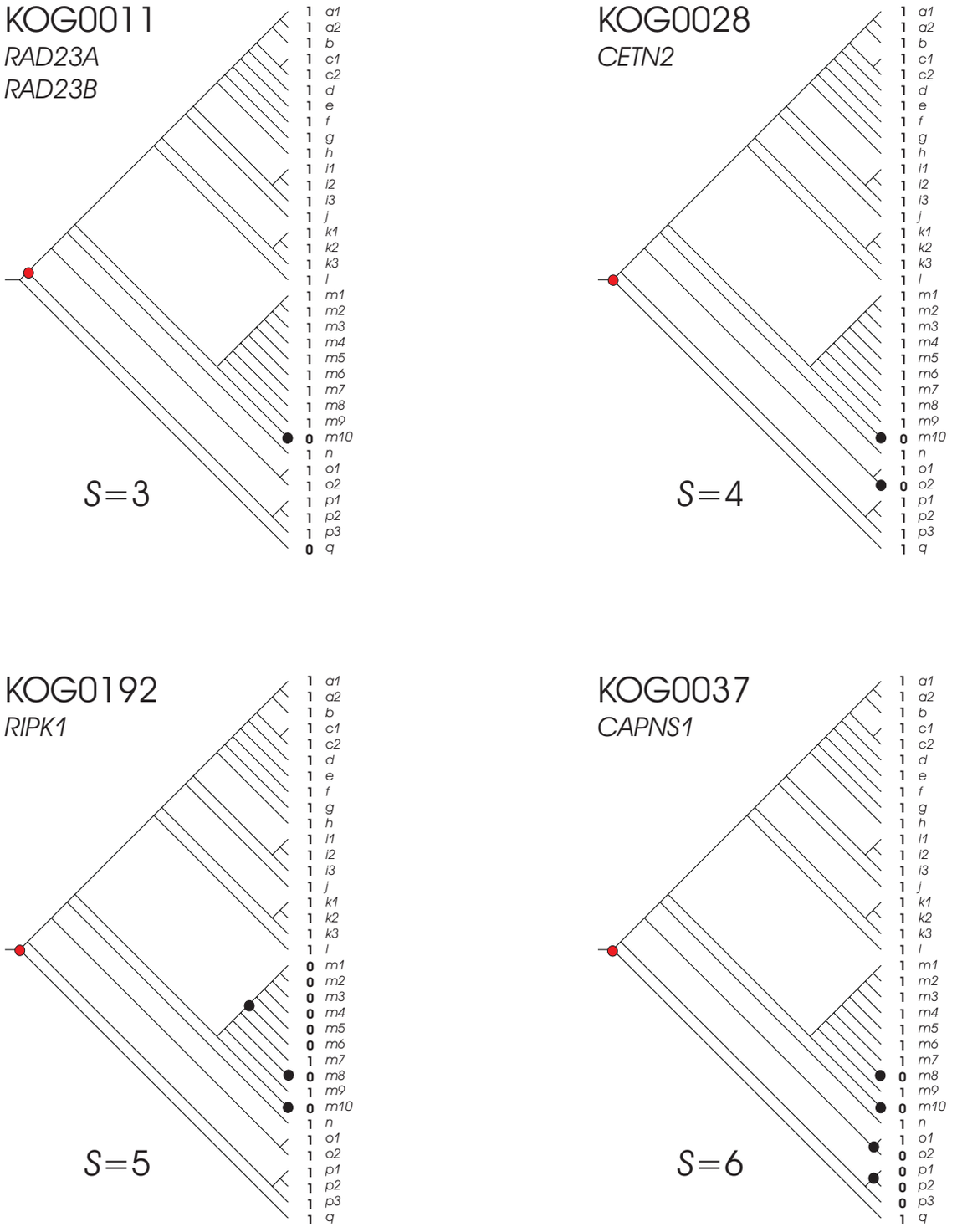
Supplementary Figure S11. Comparison of amino acid sequences of 35 eukaryotic translation initiation factor 5A proteins obtained from a three-domain orthologous groups (KOG3271/COG0231). Amino acid sequences were aligned using ClustalW. All sites containing alignment gaps and missing-information were removed before computing the distance matrix. Dots represent amino acid residues that are identical to the corresponding sites of the first sequence (*i.e. H.sapiens* sequence). Protein IDs: *H.sapiens* (ENSP00000295822); *P.troglodytes* (ENSPTRP00000026874); *M.mulatta* (ENSMMUP00000030130); *R.norvegicus* (ENSRNOP00000015779); *M.musculus* (ENSMUSP00000050289); *C.familiaris* (ENSCAFP00000022031); *B.taurus* (ENSBTAP00000002616); *M.domestica* (ENSMODP00000020009); *G.gallus* (ENSGALP00000038570); *X.tropicalis* (ENSXETP00000055749); *T.rubripes* (NEWSINFRUP00000158888); *T.nigroviridis* (GSTENP00019223001); *D.rerio* (ENSDARP00000027654); *C.ntestinalis* (ENSCINP00000012334); *D.melanogaster* (CG3186-PA); *A.gambiae* (ENSANGP00000015032); *A.mellifera* (ENSAPMP00000024661); *C.elegans* (F54C9); *K.lactis* (KLLA0E22286g); *S.cerevisiae* (YJR047C); *C.glabrata* (CAGL0L01353g); *E.gossypii* (AFR356C); *D.hansenii* (DEHA0F13640g); *Y.ipolytica* (YALI0C06886g); *A.umigatus* (Afu1g04070); *S.pombe* (SPBC336); *F.neoformans* (CND05400); *E.cuniculi* (ECU09_1370); *D.discoideum* (DDB0191442); *A.thaliana* (AT1G13950); *C.merolae* (CMS351C); *P.falciparum* (PFL0210c); *C.hominis* (Chro.70262); *T.pseudonana* (158382); *G.lamblia* (14614).



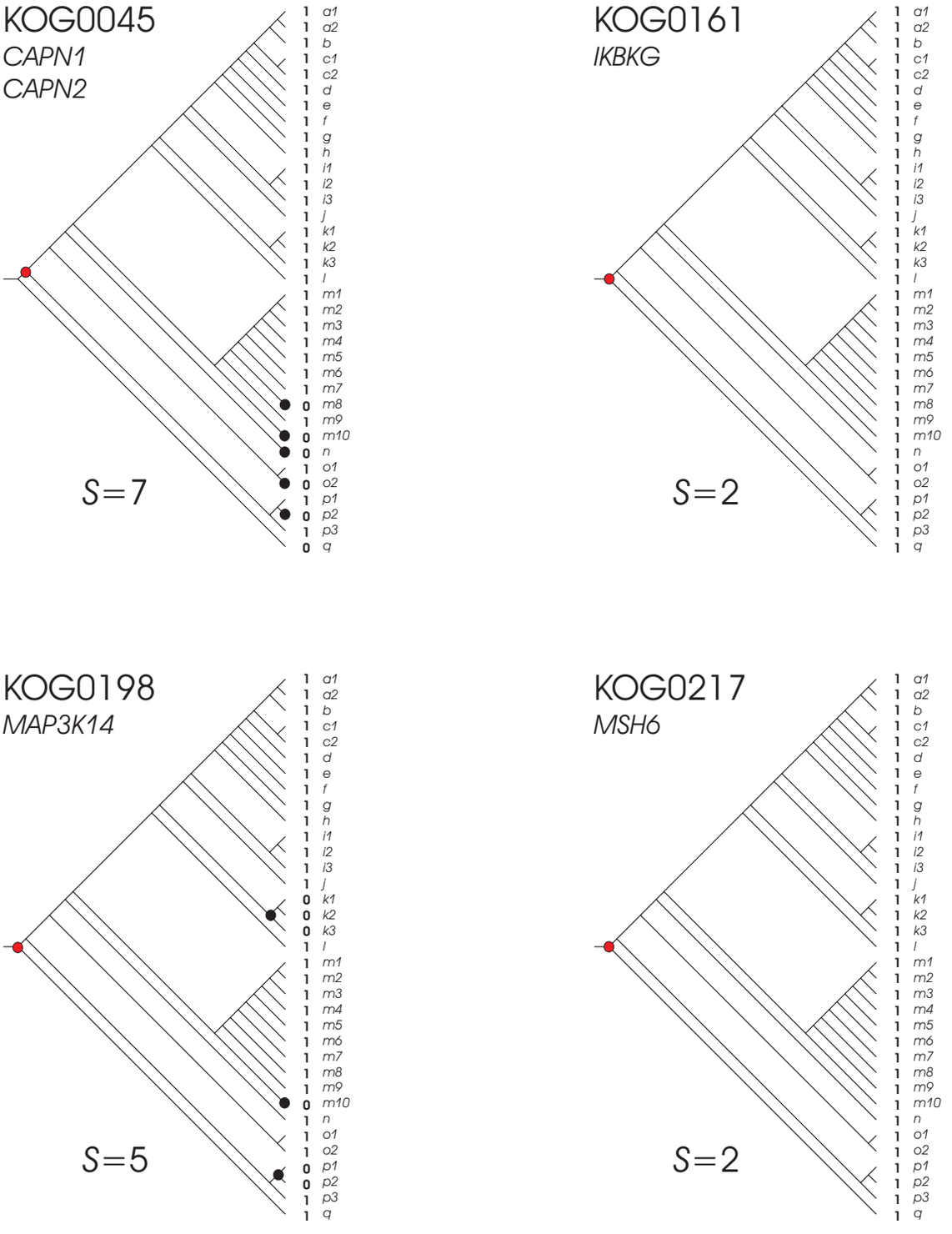
Supplementary Figure S12. Comparison of amino acid sequences of 35 eukaryotic 5-3 exonucleases (Flap Structure-specific Endonuclease 1 like proteins - FEN1) obtained from a three-domain orthologous groups (KOG2519 / COG0258). Amino acid sequences were aligned using ClustalW. All sites containing alignment gaps and missing-information were removed before computing the distance matrix. Dots represent amino acid residues that are identical to the corresponding sites of the first sequence (*i.e. H.sapiens* sequence). Protein IDs: *H.sapiens* (ENSP00000305480); *P.troglodytes* (ENSPTRP00000006454); *M.mulatta* (ENSMMUP00000008754); *R.norvegicus* (ENSRNOP00000027842); *M.musculus* (ENSMUSP00000025651); *C.familiaris* (ENSCAFP00000023634); *B.taurus* (ENSBTAP00000000071); *M.domestica* (ENSMODP00000027077); *G.gallus* (ENSGALP00000005724); *X.tropicalis* (ENSXETP00000014663); *T.rubripes* (ENSTRUP00000032042); *T.nigroviridis* (GSTENP00004156001); *D.rerio* (ENSDARP00000004016); *C.ntestinalis* (ENSCINP00000004910); *D.melanogaster* (FBpp0086223); *A.gambiae* (AGAP011448-PA); *A.mellifera* (ENSAPMP00000010885); *C.elegans* (Y47G6A.8); *K.lactis* (KLLA0F02992g); *S.cerevisiae* (YKL113C); *C.glabrata* (CAGL0K11506g); *E.gossypii* (ABL052C); *D.hansenii* (DEHA0F15059g); *Y.ipolytica* (YALI0F20042g); *A.umigatus* (Afu3g06060); *S.pombe* (SPAC3G6.06c); *F.neoformans* (CND01190); *E.cuniculi* (ECU03_1080); *D.discoideum* (DDB0186301); *A.thaliana* (AT5G26680.1); *C.merolae* (CMG106C); *P.falciparum* (PFD0420c); *C.hominis* (Chro.70245); *T.pseudonana* (132436); *G.lamblia* (16953).

Homo_sapiens																																		
Pan_troglodytes	0.00																																	
Macaca_mulatta	0.00	0.00																																
Rattus_norvegicus	0.01	0.01	0.01																															
Mus_musculus	0.01	0.01	0.01	0.01																														
Canis_familiaris	0.02	0.02	0.02	0.02	0.02																													
Bos_taurus	0.06	0.06	0.07	0.07	0.07	0.07																												
Monodelphis_domestica	0.04	0.04	0.04	0.04	0.03	0.04	0.09																											
Gallus_gallus	0.08	0.08	0.08	0.08	0.07	0.09	0.10	0.07																										
Xenopus_tropicalis	0.20	0.20	0.20	0.20	0.20	0.21	0.18	0.20	0.19																									
Takifugu_rubripes	0.16	0.16	0.16	0.16	0.16	0.18	0.14	0.16	0.15	0.19																								
Tetraodon_nigroviridis	0.16	0.16	0.16	0.17	0.16	0.18	0.14	0.16	0.15	0.19	0.02																							
Danio_rerio	0.17	0.17	0.17	0.17	0.17	0.19	0.16	0.17	0.17	0.20	0.11	0.11																						
Ciona_intestinalis	0.34	0.34	0.34	0.34	0.34	0.35	0.34	0.33	0.33	0.37	0.33	0.33	0.32																					
Drosophila_melanogaster	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.27	0.27	0.33	0.28	0.29	0.27	0.31																				
Anopheles_gambiae	0.30	0.30	0.30	0.29	0.29	0.30	0.29	0.29	0.30	0.33	0.27	0.27	0.27	0.33	0.17																			
Apis_mellifera	0.29	0.29	0.29	0.29	0.29	0.30	0.29	0.28	0.28	0.32	0.27	0.27	0.26	0.35	0.17	0.18																		
Caenorhabditis_elegans	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.37	0.29	0.29	0.31	0.40	0.32	0.33	0.32																	
Kluyveromyces_lactis	0.35	0.35	0.35	0.36	0.35	0.35	0.35	0.36	0.35	0.41	0.35	0.36	0.35	0.40	0.35	0.37	0.36	0.37																
Saccharomyces_cerevisiae	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.33	0.40	0.35	0.34	0.35	0.39	0.34	0.35	0.33	0.36	0.15															
Candida_glabrata	0.35	0.35	0.35	0.35	0.35	0.35	0.34	0.35	0.35	0.41	0.35	0.35	0.34	0.39	0.33	0.34	0.33	0.37	0.13	0.10														
Eremothecium_gossypii	0.38	0.38	0.38	0.38	0.37	0.38	0.38	0.38	0.38	0.43	0.37	0.38	0.38	0.44	0.37	0.37	0.37	0.39	0.16	0.18	0.17													
Debaryomyces_hansenii	0.34	0.34	0.34	0.34	0.34	0.34	0.33	0.34	0.34	0.39	0.34	0.34	0.33	0.37	0.33	0.32	0.34	0.35	0.20	0.19	0.18	0.24												
Yarrowia_lipolytica	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.34	0.35	0.40	0.36	0.36	0.34	0.38	0.33	0.34	0.34	0.38	0.23	0.23	0.21	0.27	0.21											
Aspergillus_fumigatus	0.36	0.36	0.36	0.36	0.36	0.37	0.38	0.35	0.36	0.41	0.38	0.38	0.38	0.40	0.36	0.37	0.37	0.38	0.29	0.28	0.28	0.32	0.27	0.27										
Schizosaccharomyces_pombe	0.35	0.35	0.35	0.36	0.35	0.36	0.35	0.34	0.34	0.41	0.37	0.37	0.37	0.39	0.37	0.39	0.38	0.38	0.32	0.31	0.32	0.36	0.32	0.32	0.32									
Filobasidiella_neoformans	0.39	0.39	0.38	0.38	0.38	0.38	0.39	0.38	0.38	0.43	0.38	0.38	0.38	0.43	0.38	0.39	0.37	0.41	0.30	0.30	0.29	0.34	0.30	0.30	0.31	0.35								
Encephalitozoon_cuniculi	0.62	0.62	0.62	0.62	0.62	0.62	0.61	0.61	0.61	0.64	0.62	0.62	0.60	0.63	0.60	0.61	0.62	0.64	0.63	0.63	0.62	0.65	0.62	0.61	0.63	0.63	0.63							
Dictyostelium_discoideum	0.37	0.37	0.37	0.37	0.37	0.38	0.36	0.36	0.36	0.43	0.37	0.38	0.38	0.45	0.41	0.40	0.39	0.42	0.43	0.43	0.43	0.45	0.42	0.39	0.39	0.41	0.42	0.62						
Arabidopsis_thaliana	0.36	0.36	0.36	0.36	0.36	0.36	0.37	0.34	0.34	0.41	0.36	0.35	0.35	0.40	0.36	0.34	0.36	0.38	0.37	0.36	0.36	0.41	0.35	0.35	0.37	0.39	0.39	0.59	0.39					
Cyanidioschyzon_merolae	0.45	0.45	0.45	0.45	0.45	0.45	0.46	0.44	0.45	0.50	0.45	0.44	0.44	0.50	0.46	0.47	0.44	0.48	0.48	0.47	0.47	0.51	0.47	0.46	0.46	0.50	0.48	0.67	0.50	0.42				
Plasmodium_falciparum	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.40	0.47	0.41	0.42	0.40	0.45	0.39	0.38	0.40	0.41	0.43	0.43	0.41	0.46	0.43	0.41	0.44	0.45	0.45	0.63	0.42	0.37	0.44			
Cryptosporidium_hominis	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.43	0.43	0.48	0.42	0.42	0.42	0.45	0.42	0.42	0.42	0.44	0.44	0.43	0.44	0.48	0.45	0.42	0.42	0.47	0.49	0.64	0.45	0.39	0.45	0.34		
Thalassiosira_pseudonana	0.42	0.42	0.42	0.43	0.42	0.43	0.43	0.42	0.43	0.47	0.45	0.45	0.43	0.44	0.43	0.43	0.44	0.44	0.43	0.44	0.43	0.47	0.43	0.41	0.42	0.44	0.44	0.63	0.42	0.37	0.49	0.42	0.44	
Giardia_lamblia	0.49	0.49	0.49	0.49	0.49	0.48	0.49	0.49	0.48	0.54	0.49	0.49	0.51	0.53	0.49	0.49	0.48	0.48	0.53	0.51	0.51	0.54	0.53	0.52	0.53	0.52	0.54	0.66	0.49	0.47	0.51	0.51	0.52	0.50

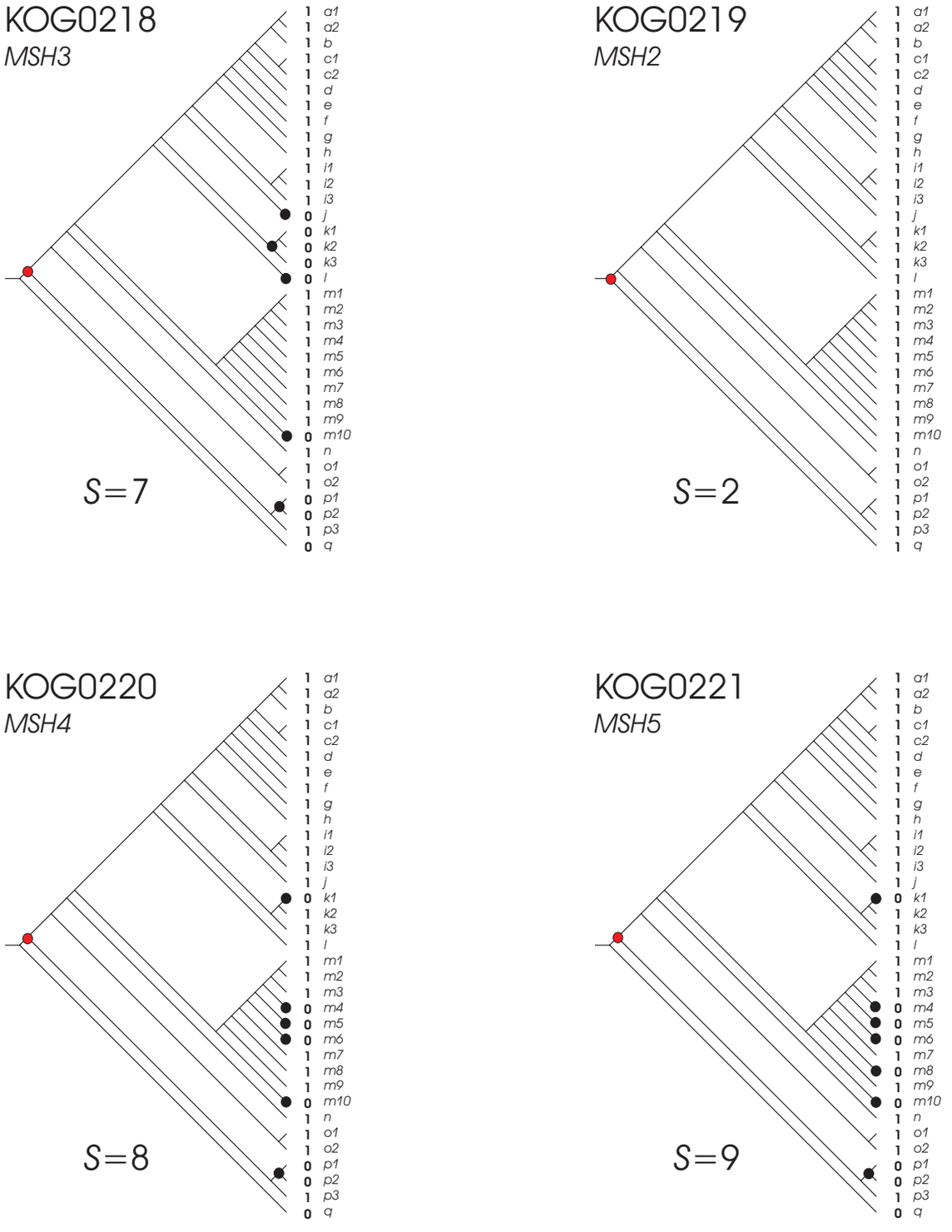
Supplementary Figure S13. Divergence matrix according to the alignments presented in **Supplementary Figures S10-S12**. The number of amino acid differences per site from analysis between sequences is shown. All results are based on the pairwise analysis of 35 sequences (average of the three proteins families). Analyses were conducted in MEGA4 (Tamura K. *et al*., *Mol. Biol. Evol.* **24**:1596-1599, 2007). All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option).



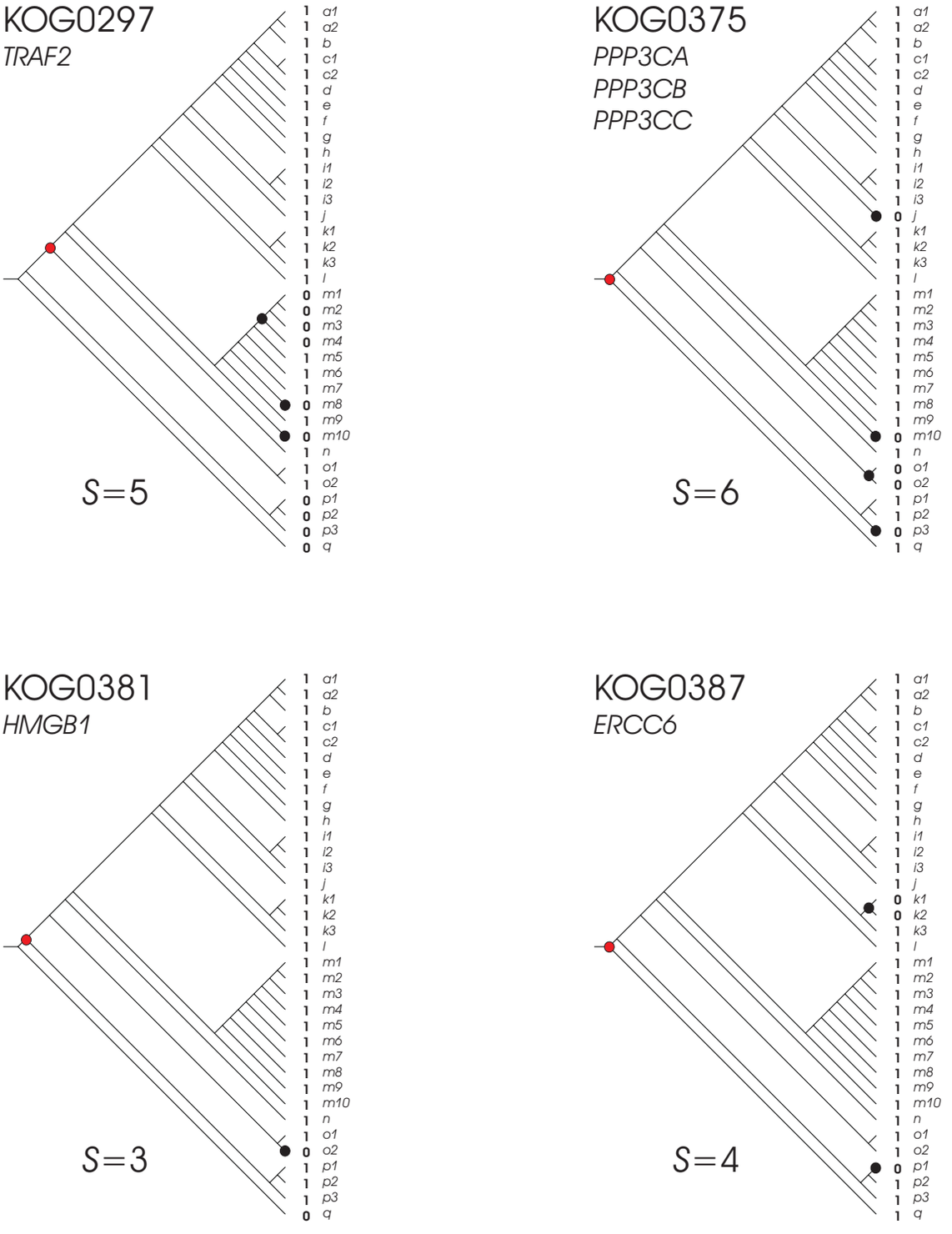
Supplementary Figure S14. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0011, KOG0028, KOG0192 and KOG0037. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes: ***a1*** (*Homo sapiens*); ***a2*** (*Pan troglodytes*); ***b*** (*Macaca mulatta*); ***c1*** (*Rattus norvegicus*); ***c2*** (*Mus musculus*); ***d*** (*Canis familiaris*); ***e1*** (*Bos Taurus*); ***f*** (*Monodelphis domestica*); ***g*** (*Gallus gallus*); ***h*** (*Xenopus tropicalis*); ***i1*** (*Takifugu rubripes*); ***i2*** (*Tetraodon nigroviridis*); ***i3*** (*Danio rerio*); ***j*** (*Ciona intestinalis*); ***k1*** (*Drosophila melanogaster*); ***k2*** (*Anopheles gambiae*); ***k3*** (*Apis mellifera*); ***l*** (*Caenorhabditis elegans*); ***m1*** (*Kluyveromyces lactis*); ***m2*** (*Saccharomyces cerevisiae*); ***m3*** (*Candida glabrata*); ***m4*** (*Eremothecium gossypii*); ***m5*** (*Debaryomyces hansenii*); ***m6*** (*Yarrowia lipolytica*); ***m7*** (*Aspergillus fumigatus*); ***m8*** (*Schizosaccharomyces pombe*); ***m9*** (*Filobasidiella neoformans*); ***m10*** (*Encephalitozoon cuniculi*); ***n*** (*Dictyostelium discoideum*); ***o1*** (*Arabidopsis thaliana*); ***o2*** (*Cyanidioschyzon merolae*); ***p1*** (*Plasmodium falciparum*); ***p2*** (*Cryptosporidium hominis*); ***p3*** (*Thalassiosira pseudonana CCMP1335*); ***q*** (*Giardia lamblia ATCC 50803*).



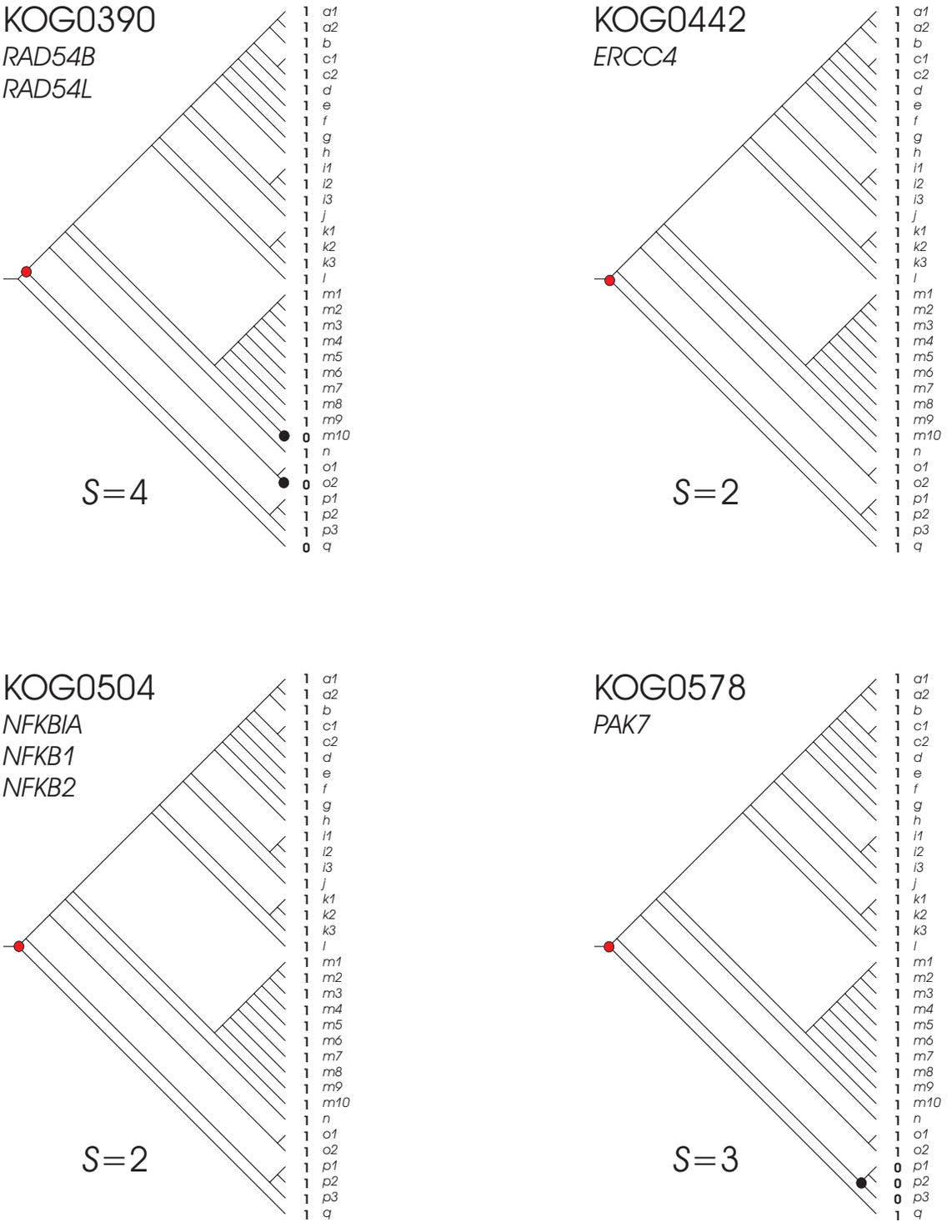
Supplementary Figure S15. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0045, KOG0161, KOG0198 and KOG0217, as in Supplementary Figure S14.



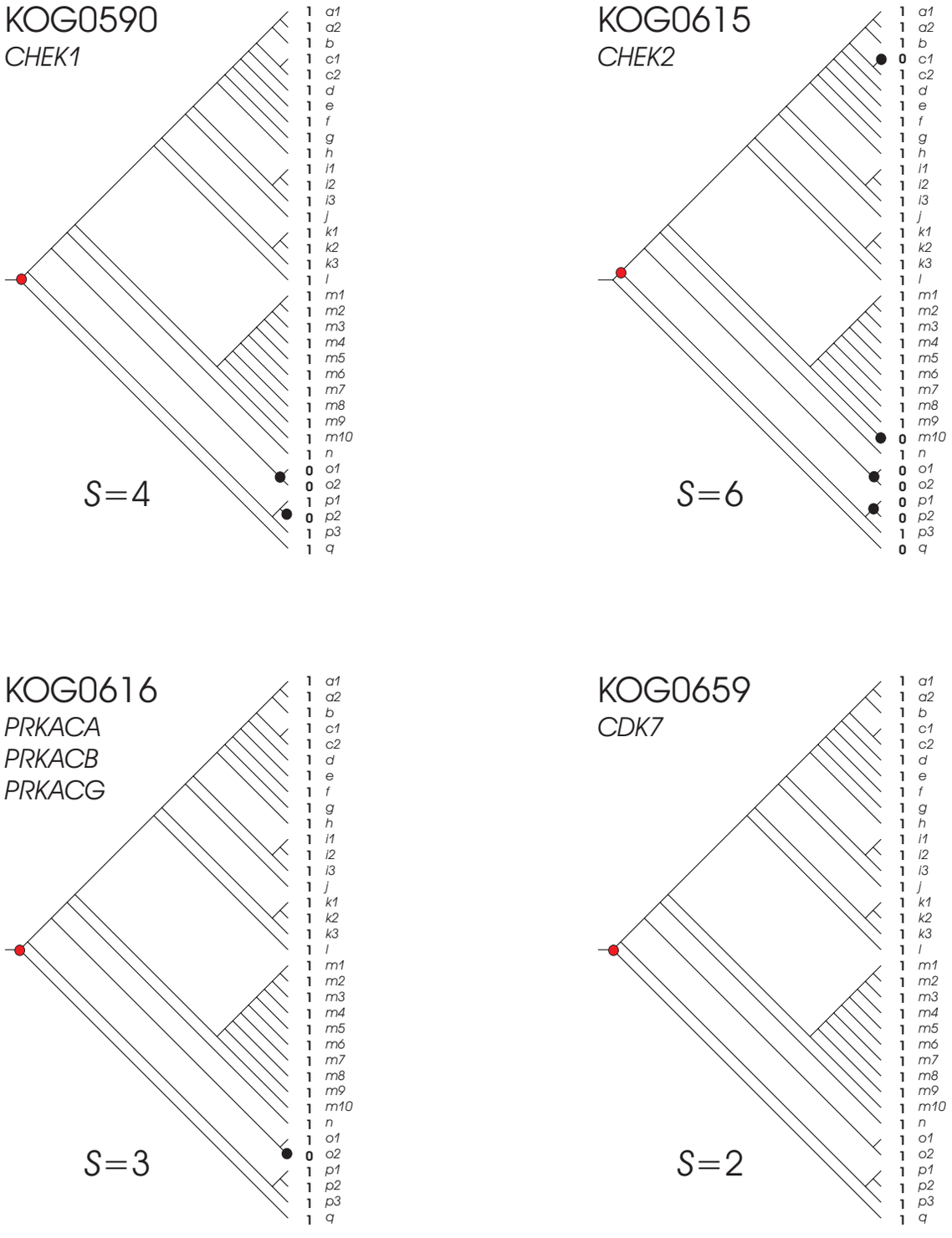
Supplementary Figure S16. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0218, KOG0219, KOG0220 and KOG0221, as in Supplementary Figure S14.



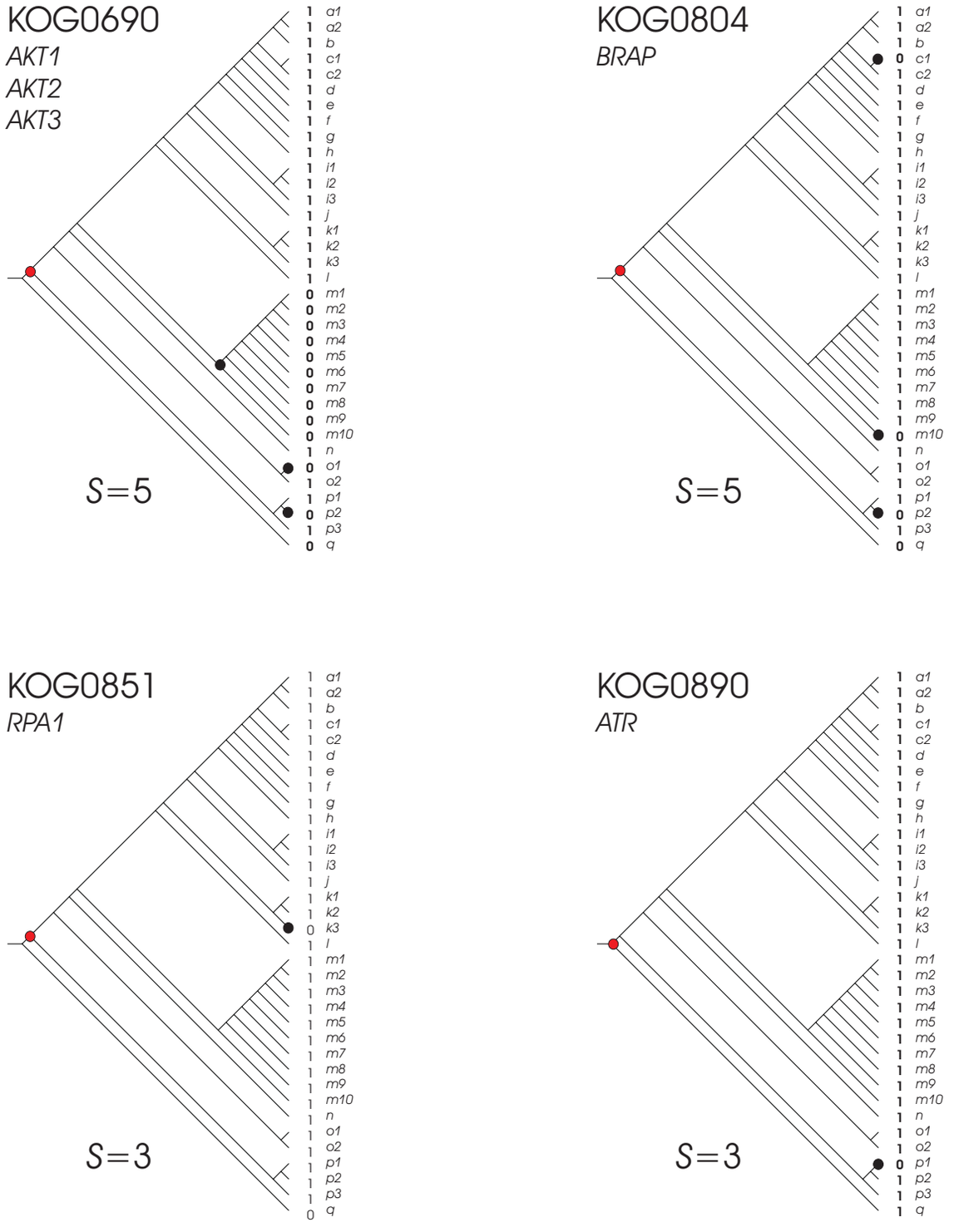
Supplementary Figure S17. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0297, KOG0375, KOG0381 and KOG0387, as in Supplementary Figure S14.



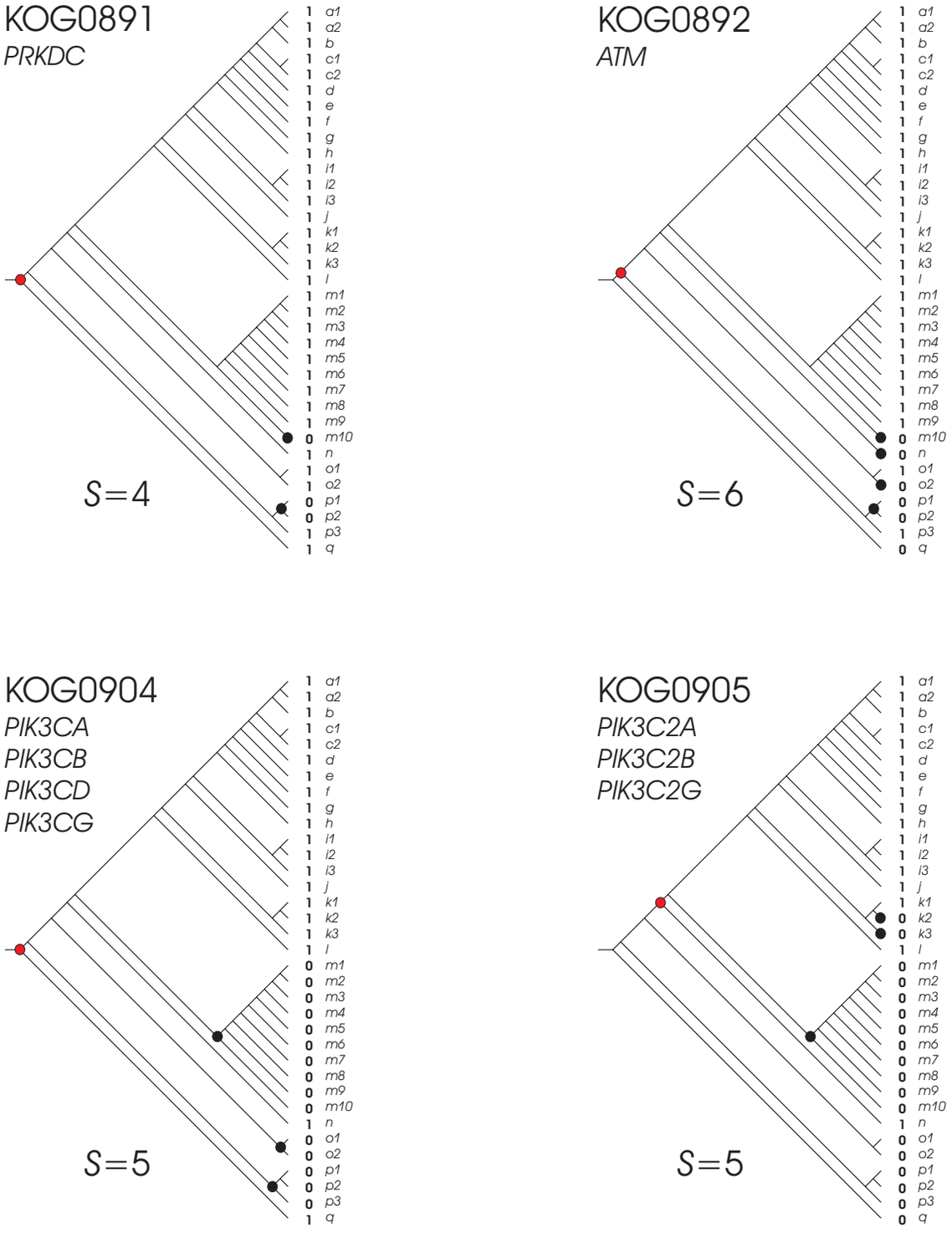
Supplementary Figure S18. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0390, KOG0442, KOG0504 and KOG0578, as in Supplementary Figure S14.



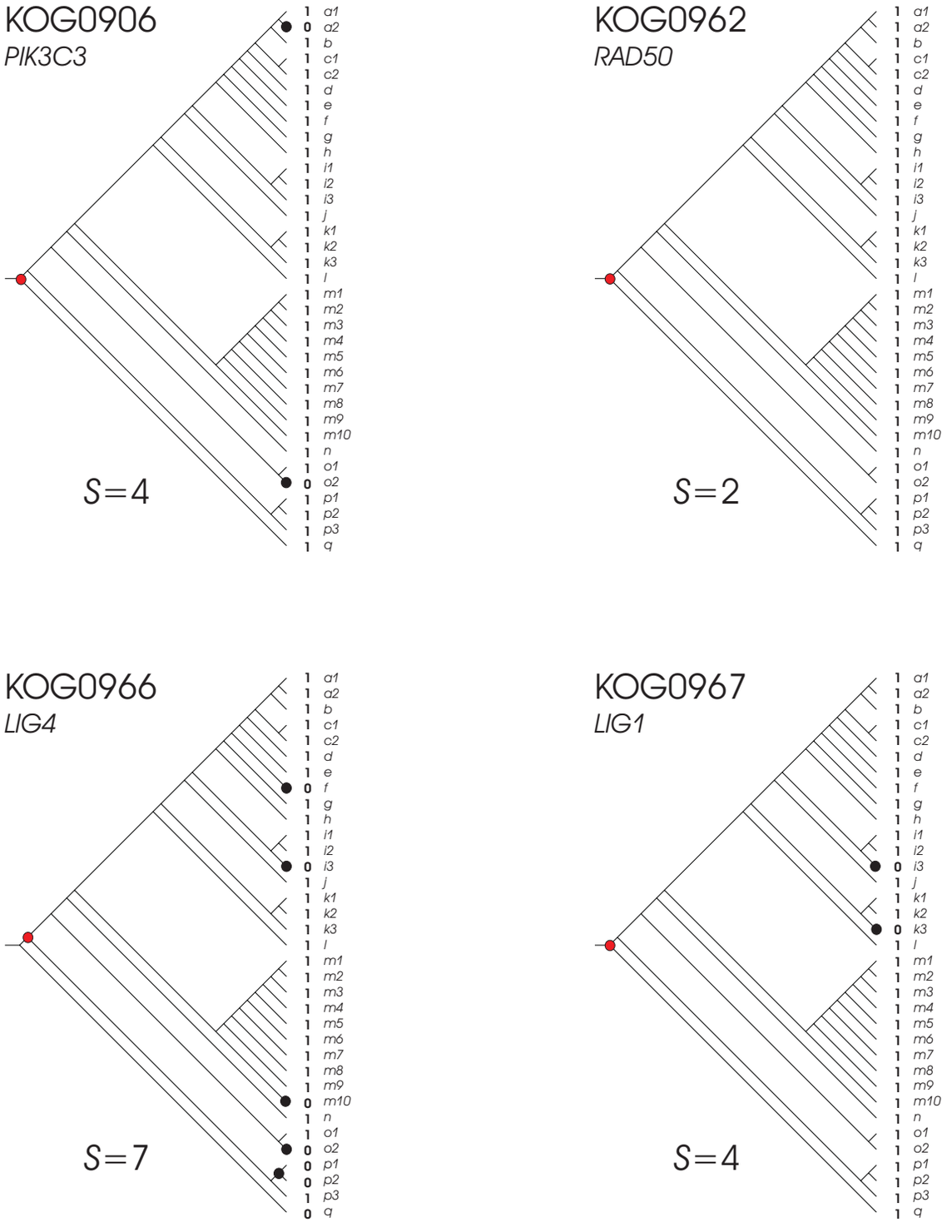
Supplementary Figure S19. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0590, KOG0615, KOG0616 and KOG0659, as in Supplementary Figure S14.



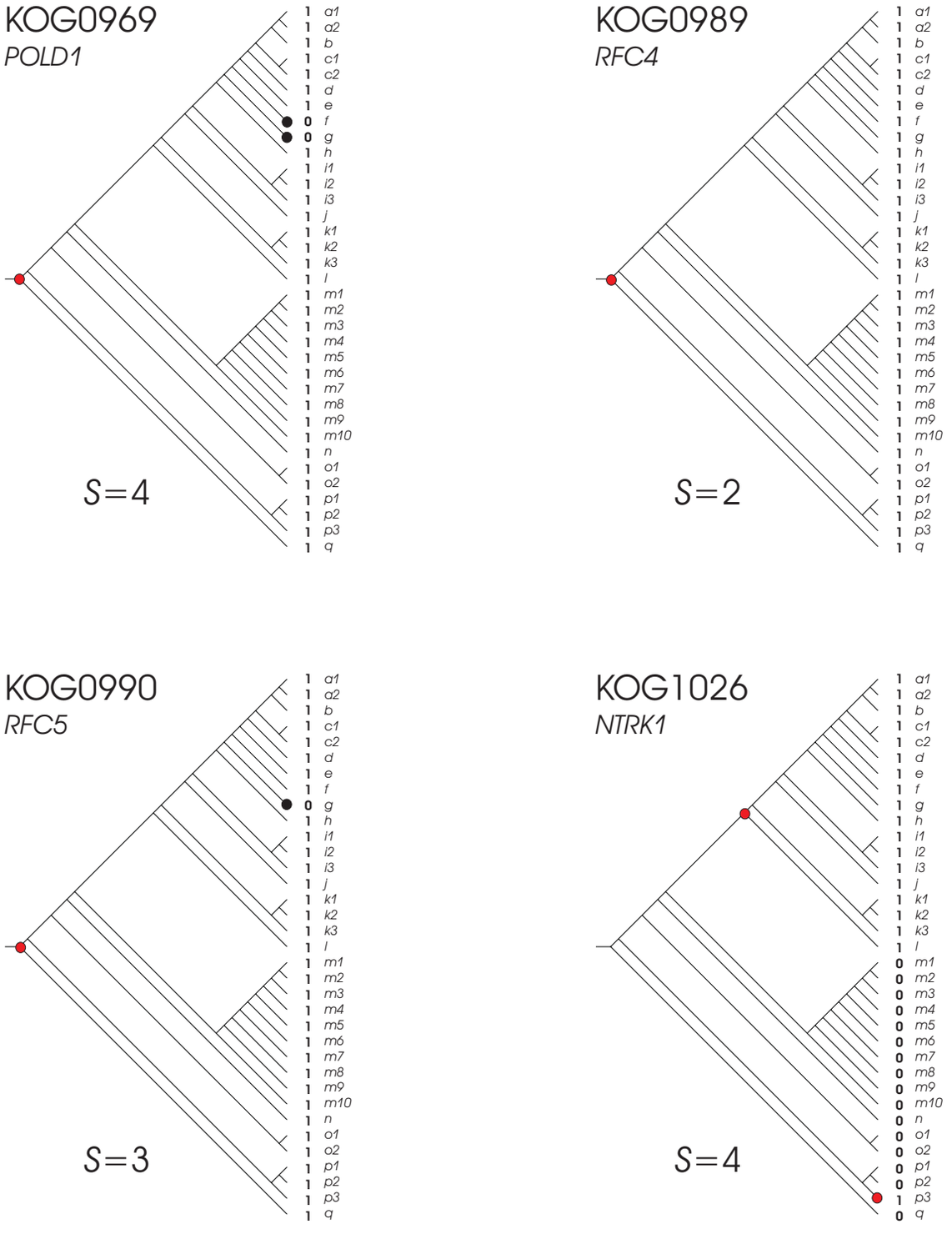
Supplementary Figure S20. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0690, KOG0804, KOG0851 and KOG0890, as in Supplementary Figure S14.



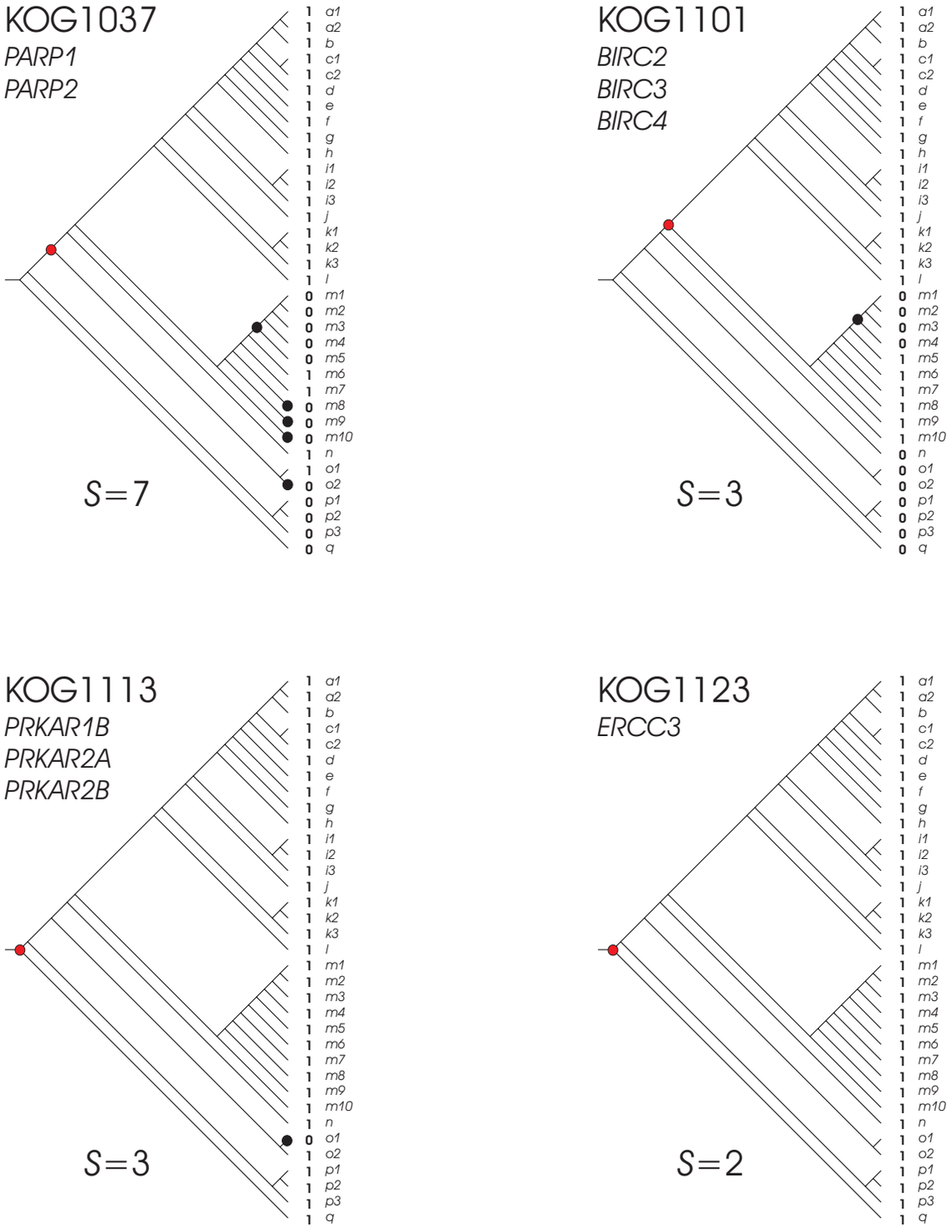
Supplementary Figure S21. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0891, KOG0892, KOG0904 and KOG0905, as in Supplementary Figure S14.



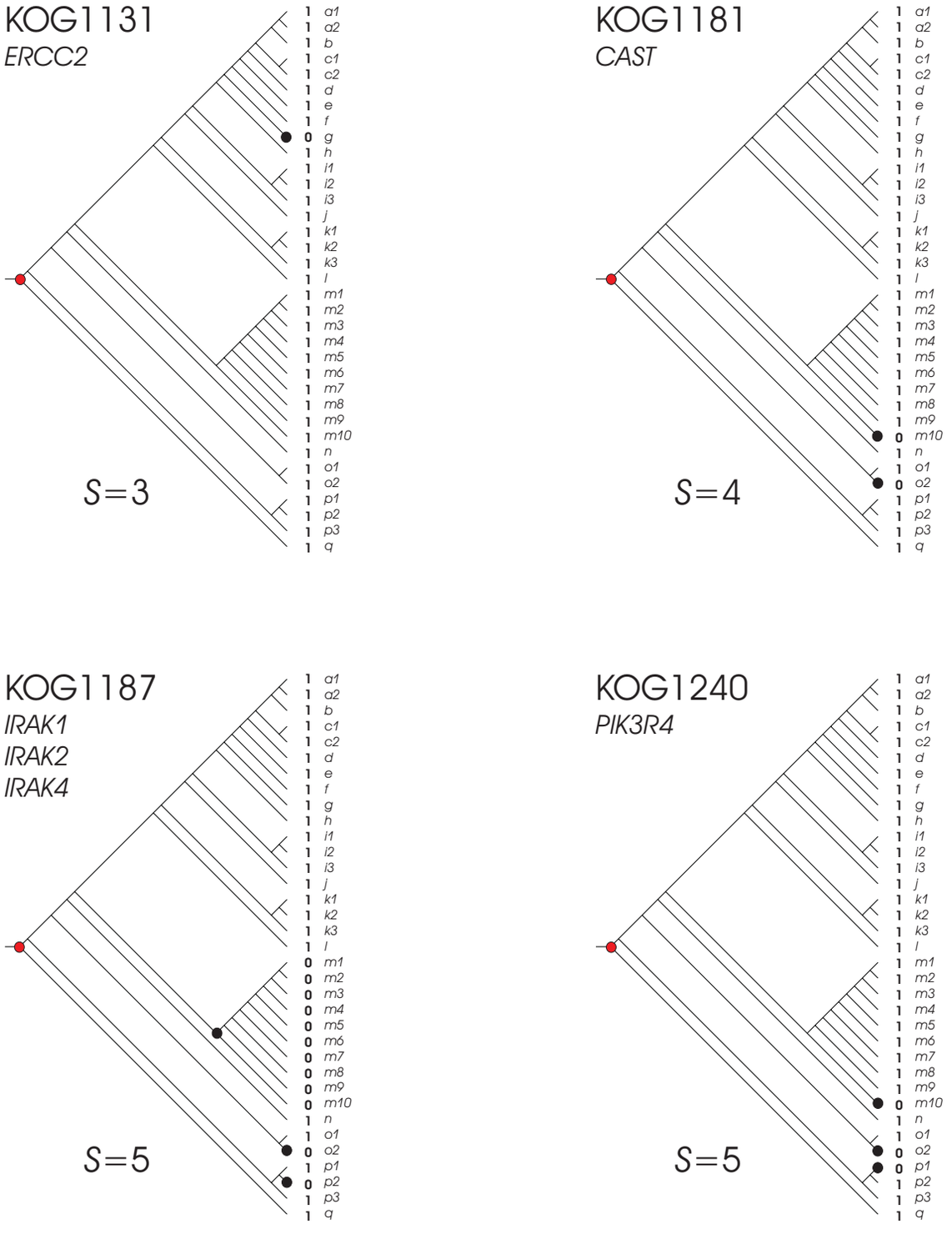
Supplementary Figure S22. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0906, KOG0962, KOG0966 and KOG0967, as in Supplementary Figure S14.



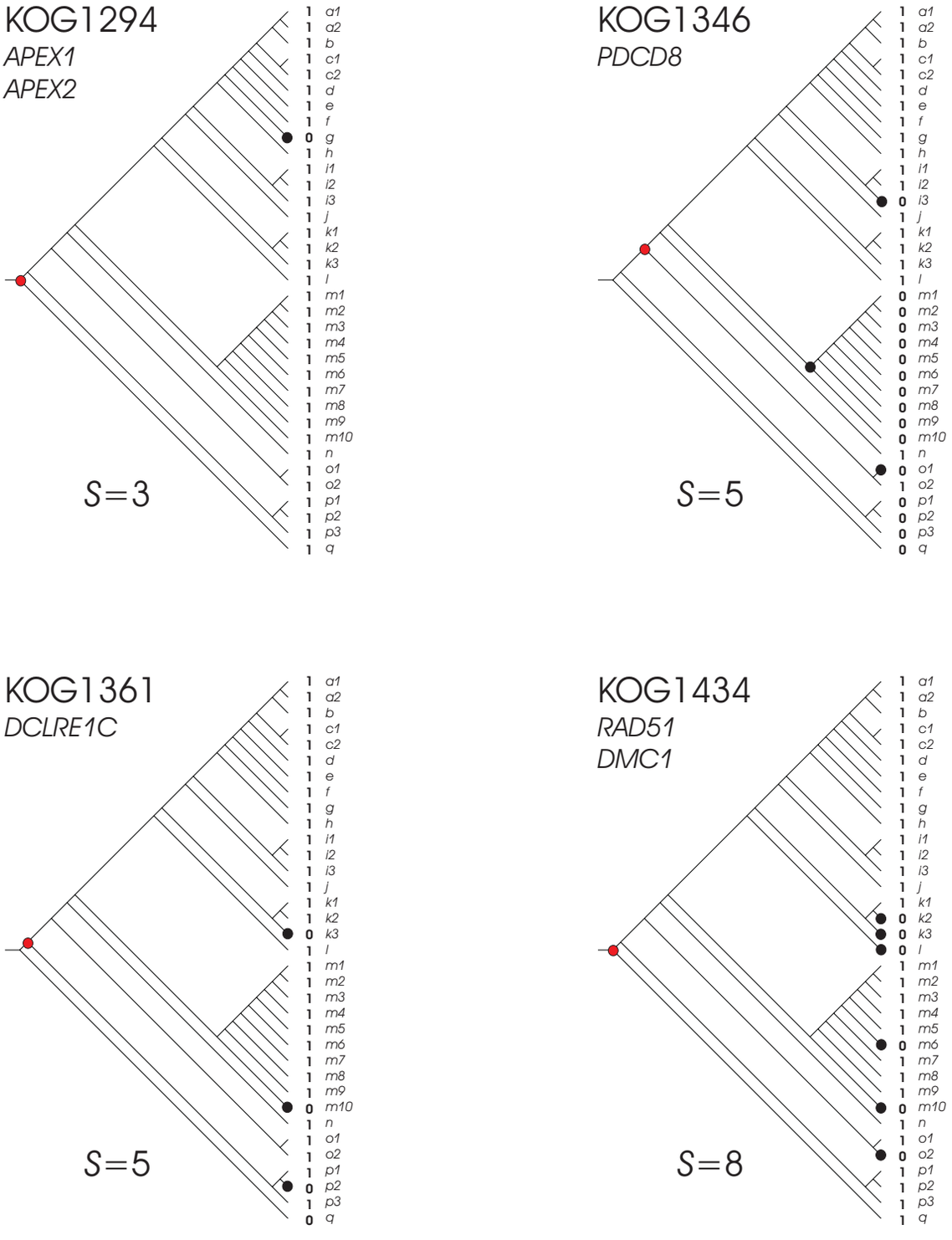
Supplementary Figure S23. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0969, KOG0989, KOG0990 and KOG1026, as in Supplementary Figure S14.



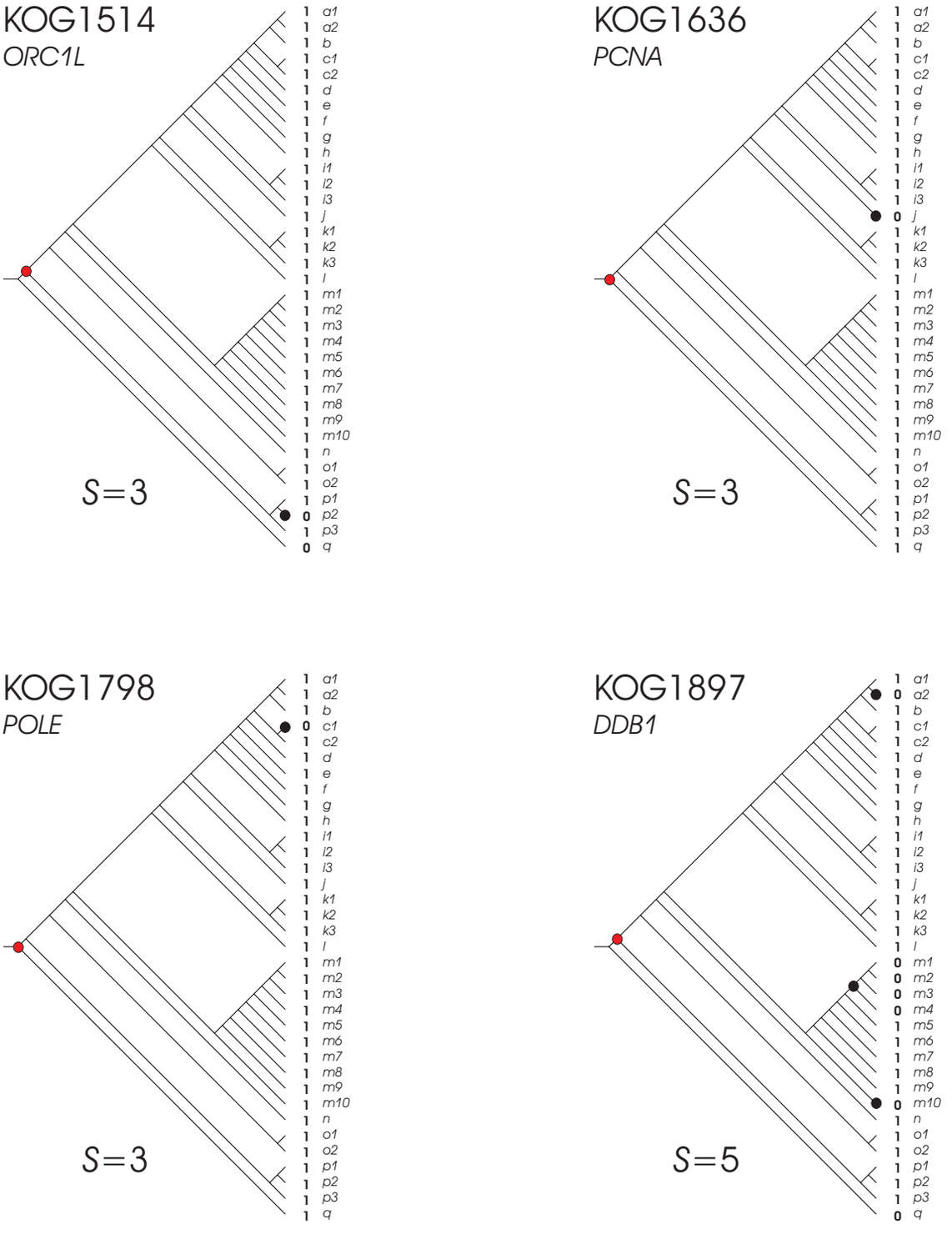
Supplementary Figure S24. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1037, KOG1101, KOG1113 and KOG1123, as in Supplementary Figure S14.



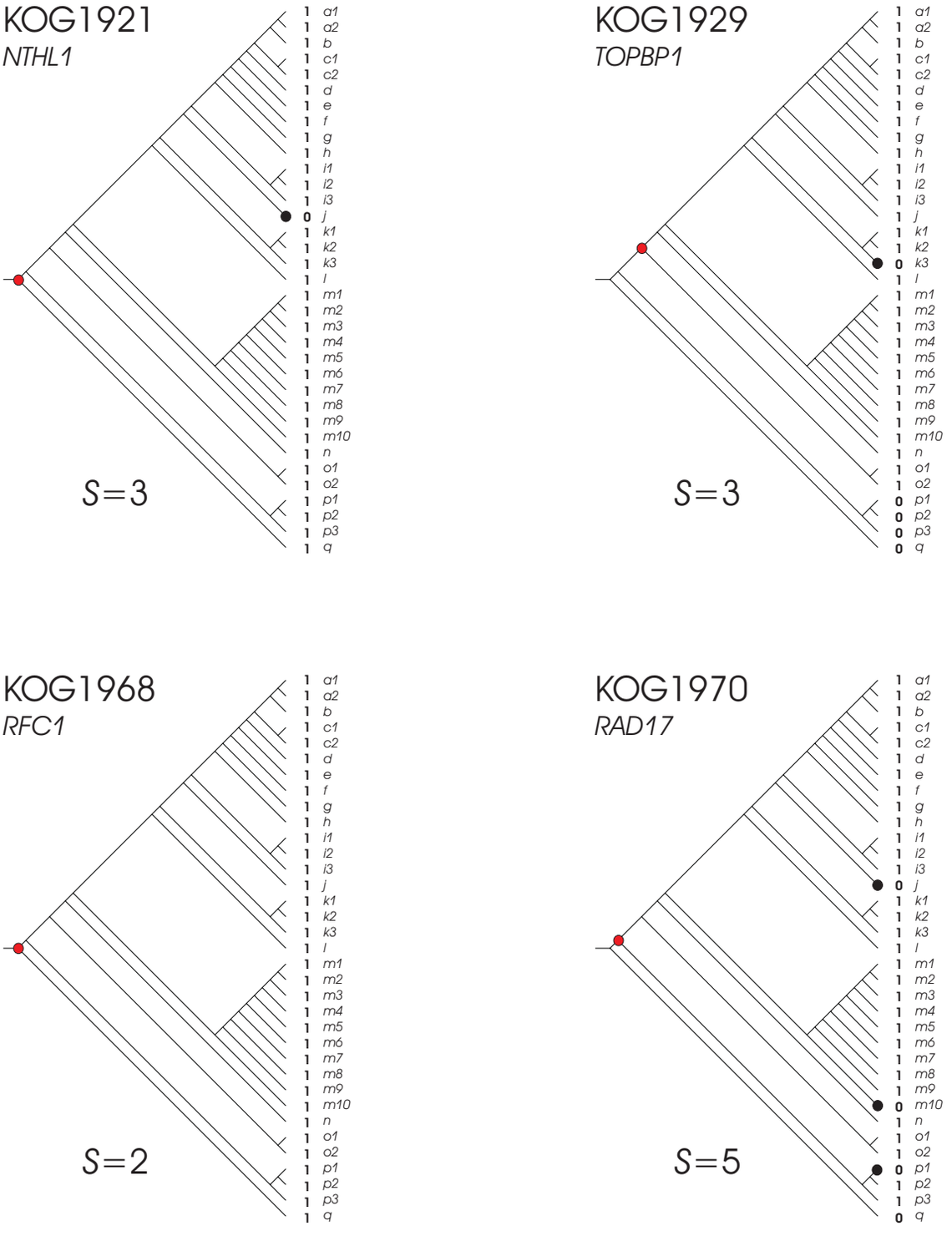
Supplementary Figure S25. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1131, KOG1181, KOG1187 and KOG1240, as in Supplementary Figure S14.



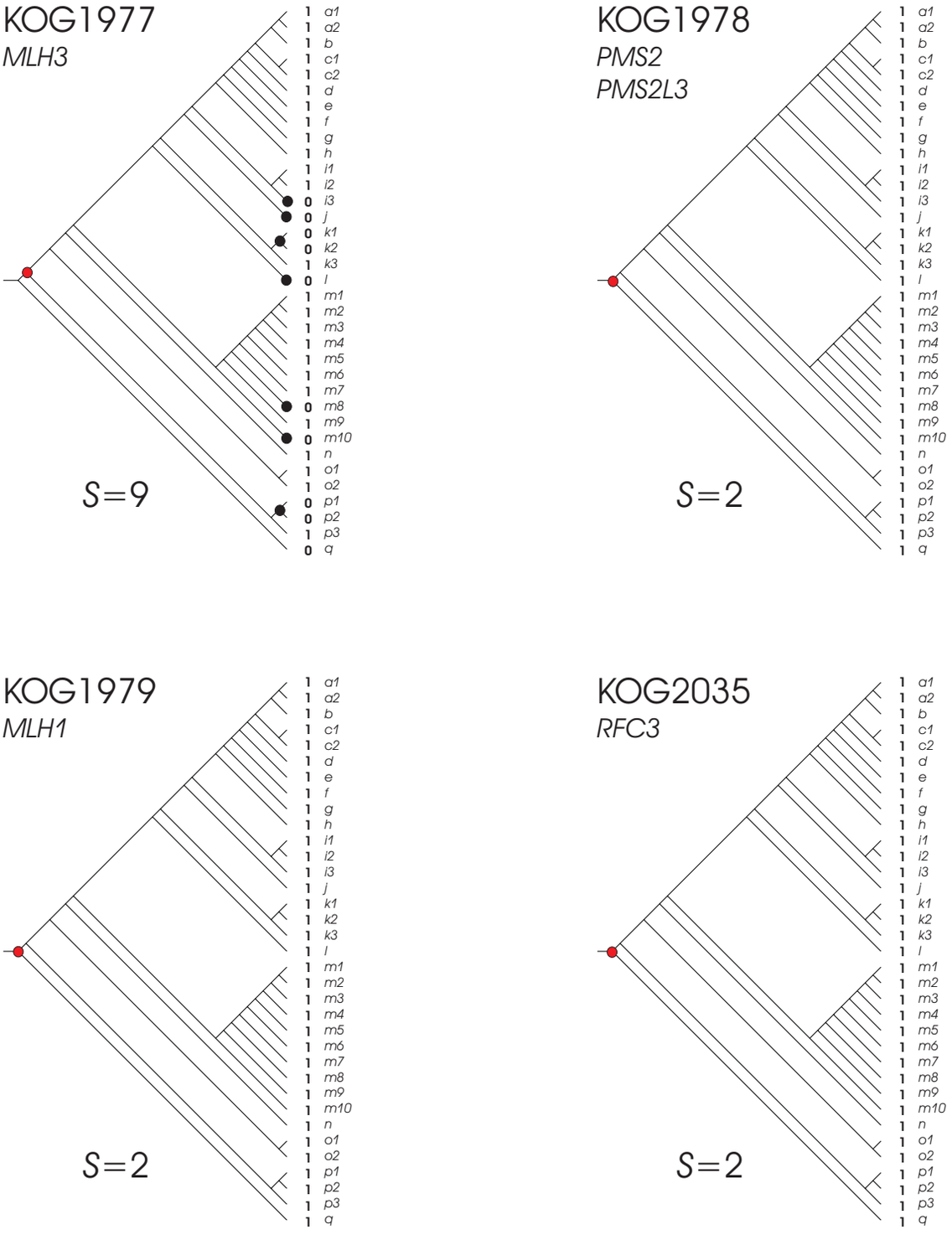
Supplementary Figure S26. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1294, KOG1346, KOG1361 and KOG1434, as in Supplementary Figure S14.



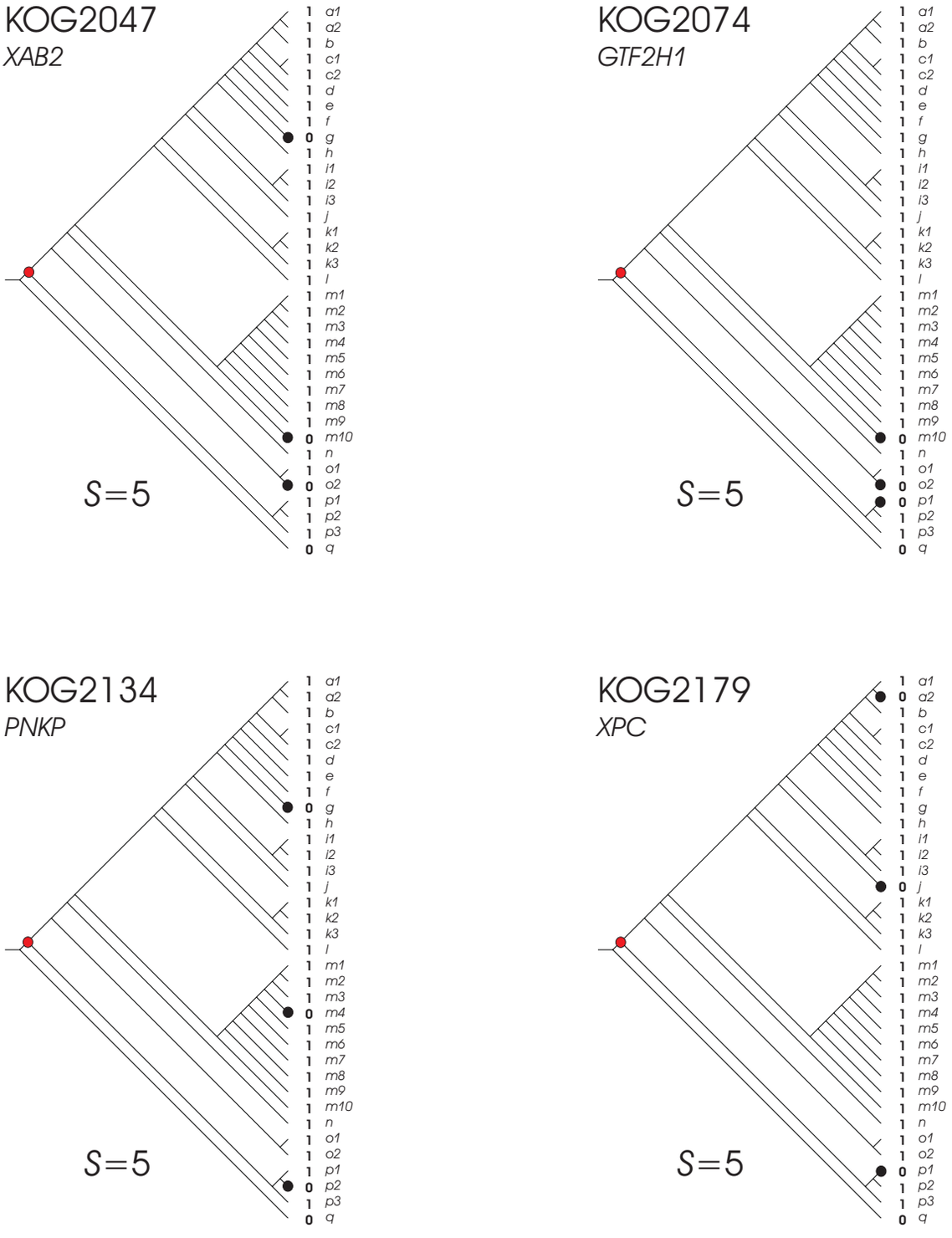
Supplementary Figure S27. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1514, KOG1636, KOG1798 and KOG1897, as in Supplementary Figure S14.



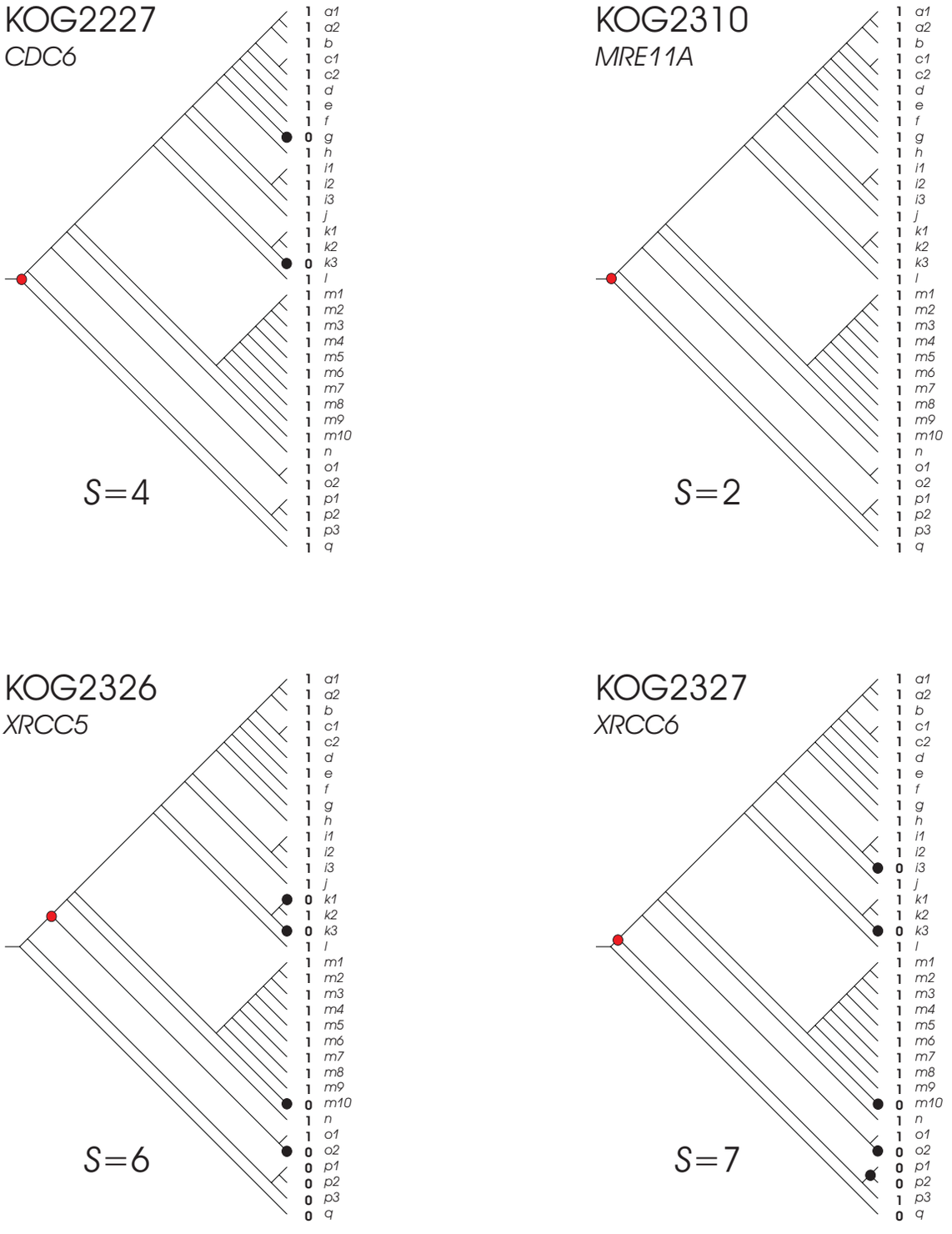
Supplementary Figure S28. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1921, KOG1929, KOG1968 and KOG1970, as in Supplementary Figure S14.



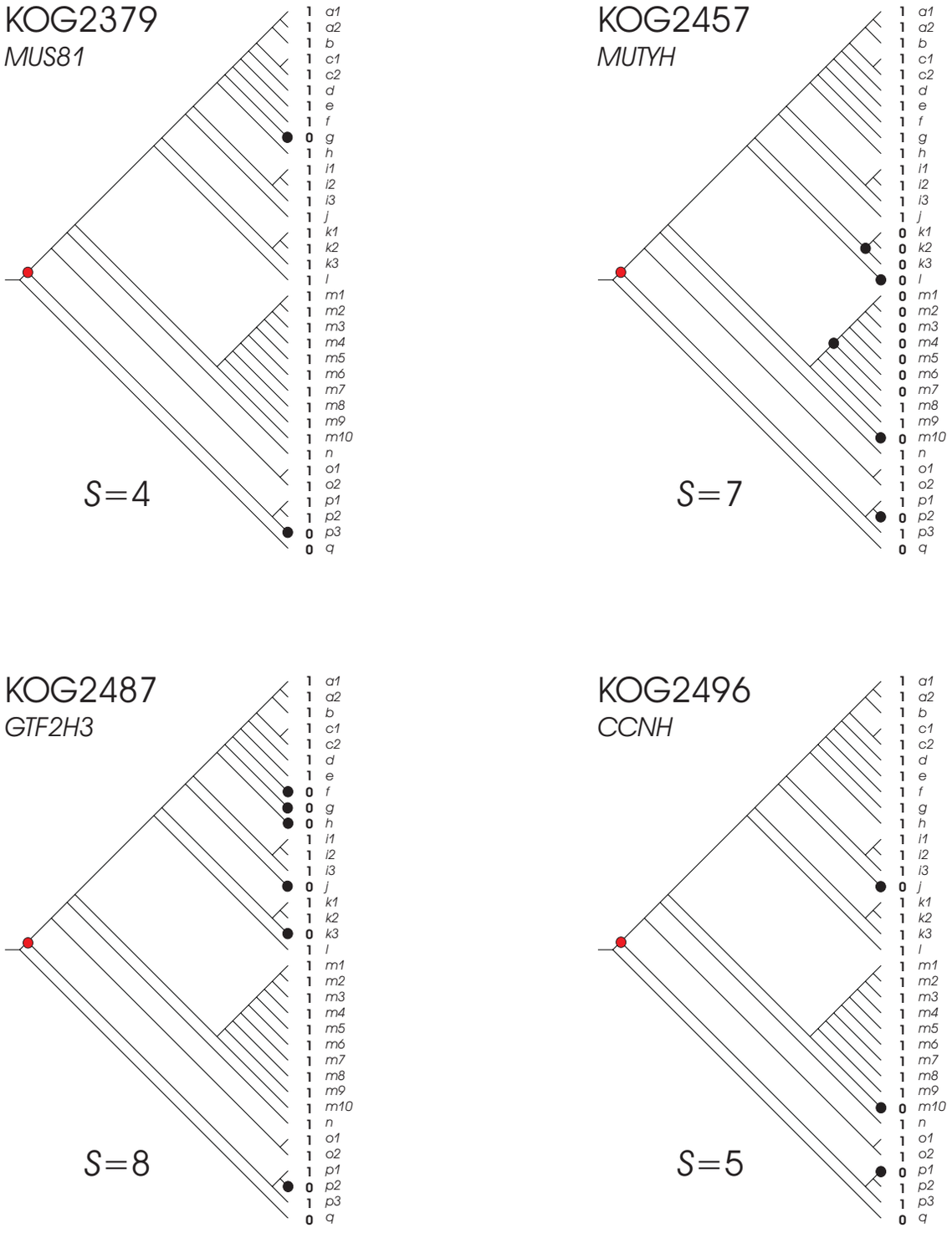
Supplementary Figure S29. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1977, KOG1978, KOG1979 and KOG2035, as in Supplementary Figure S14.



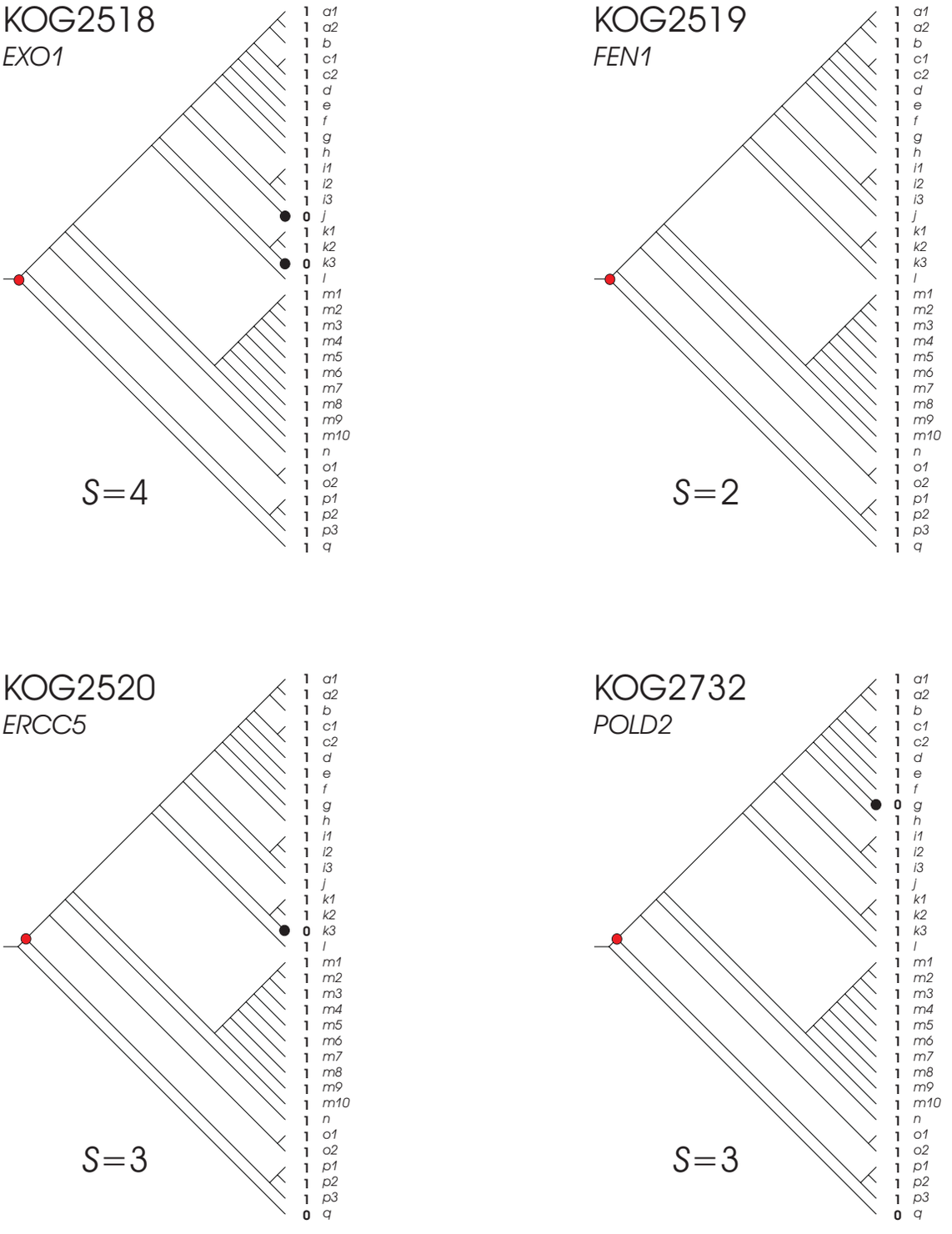
Supplementary Figure S30. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2047, KOG2074, KOG2134 and KOG2179, as in Supplementary Figure S14.



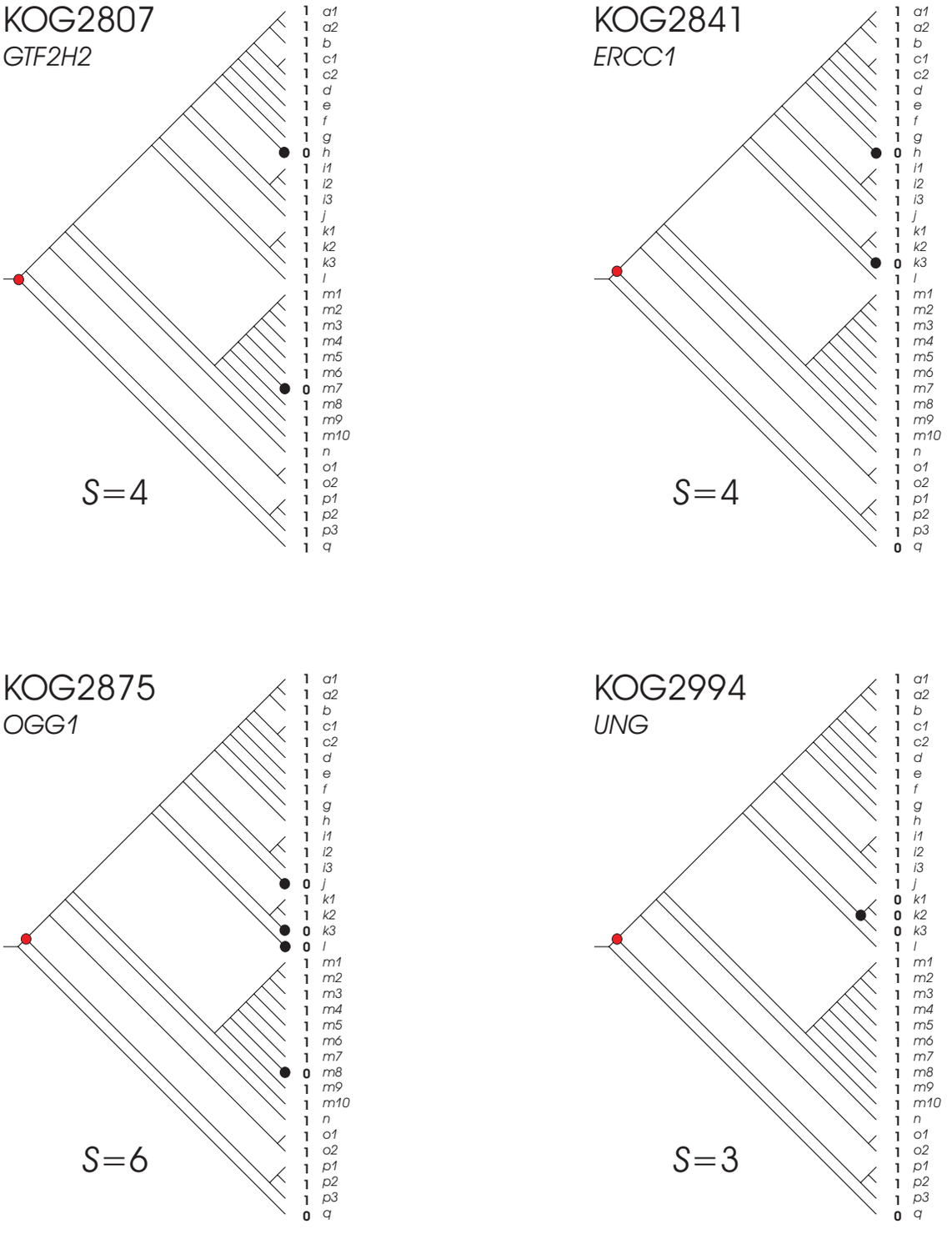
Supplementary Figure S31. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2227, KOG2310, KOG2326 and KOG2327, as in Supplementary Figure S14.



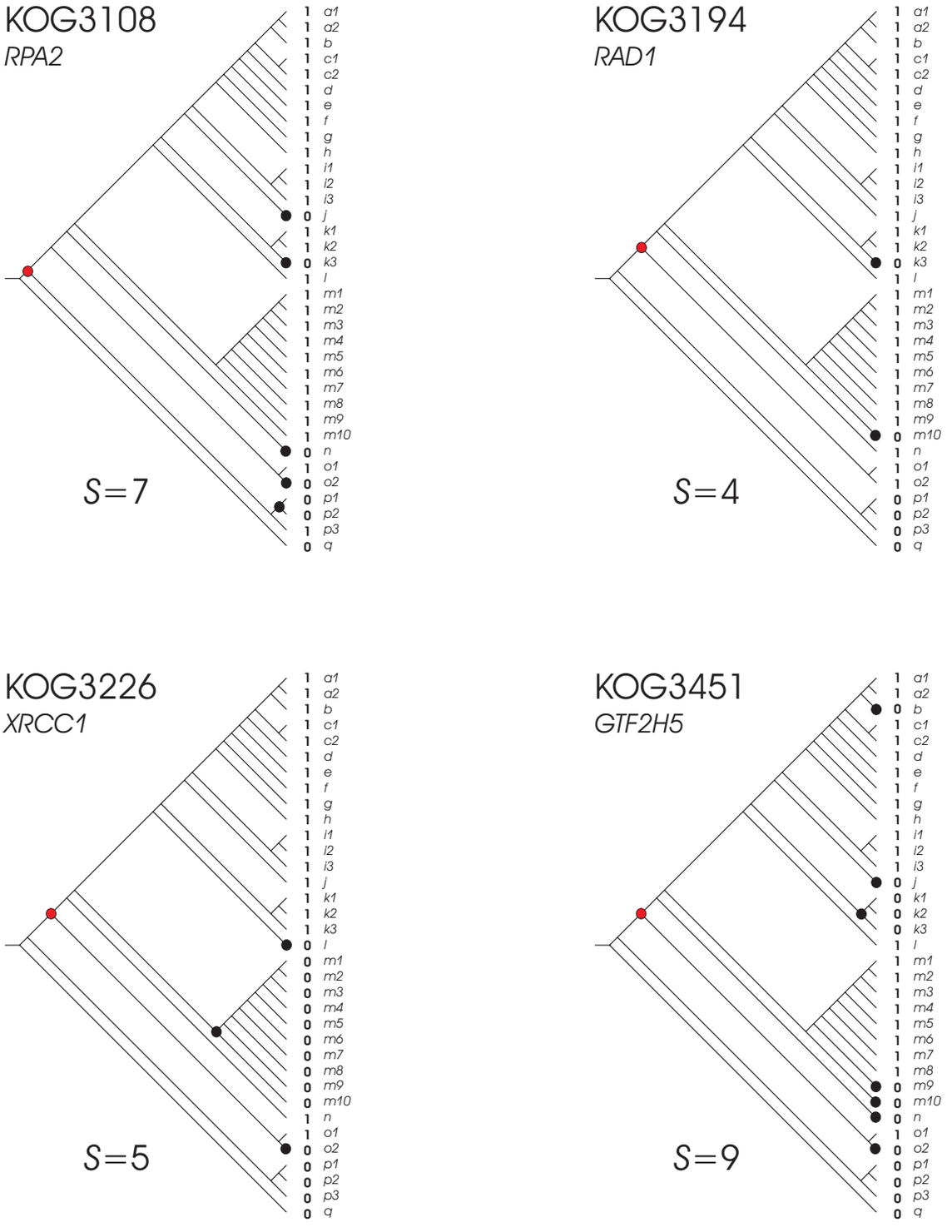
Supplementary Figure S32. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2379, KOG2457, KOG2487 and KOG2496, as in Supplementary Figure S14.



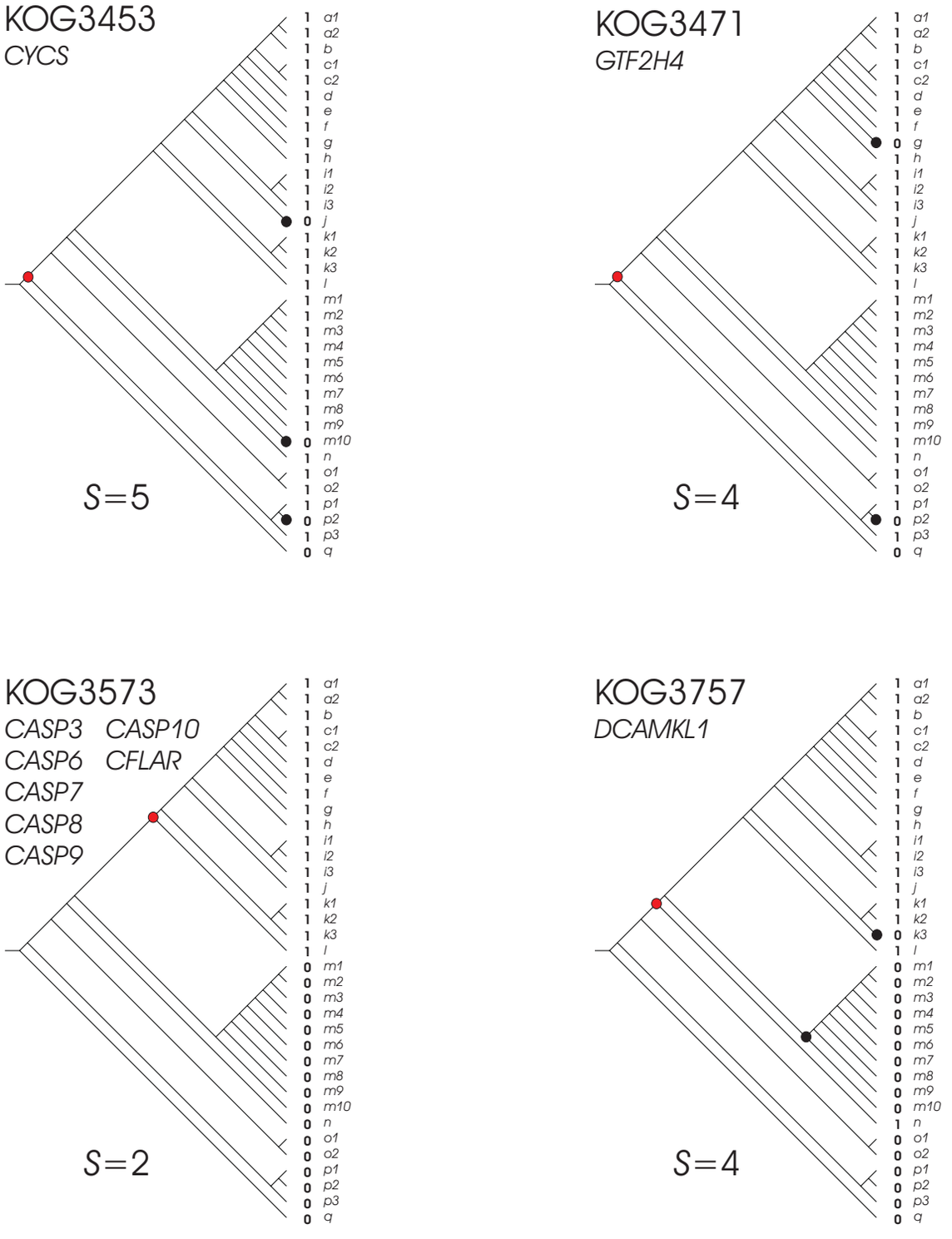
Supplementary Figure S33. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2518, KOG2519, KOG2520 and KOG2732, as in Supplementary Figure S14.



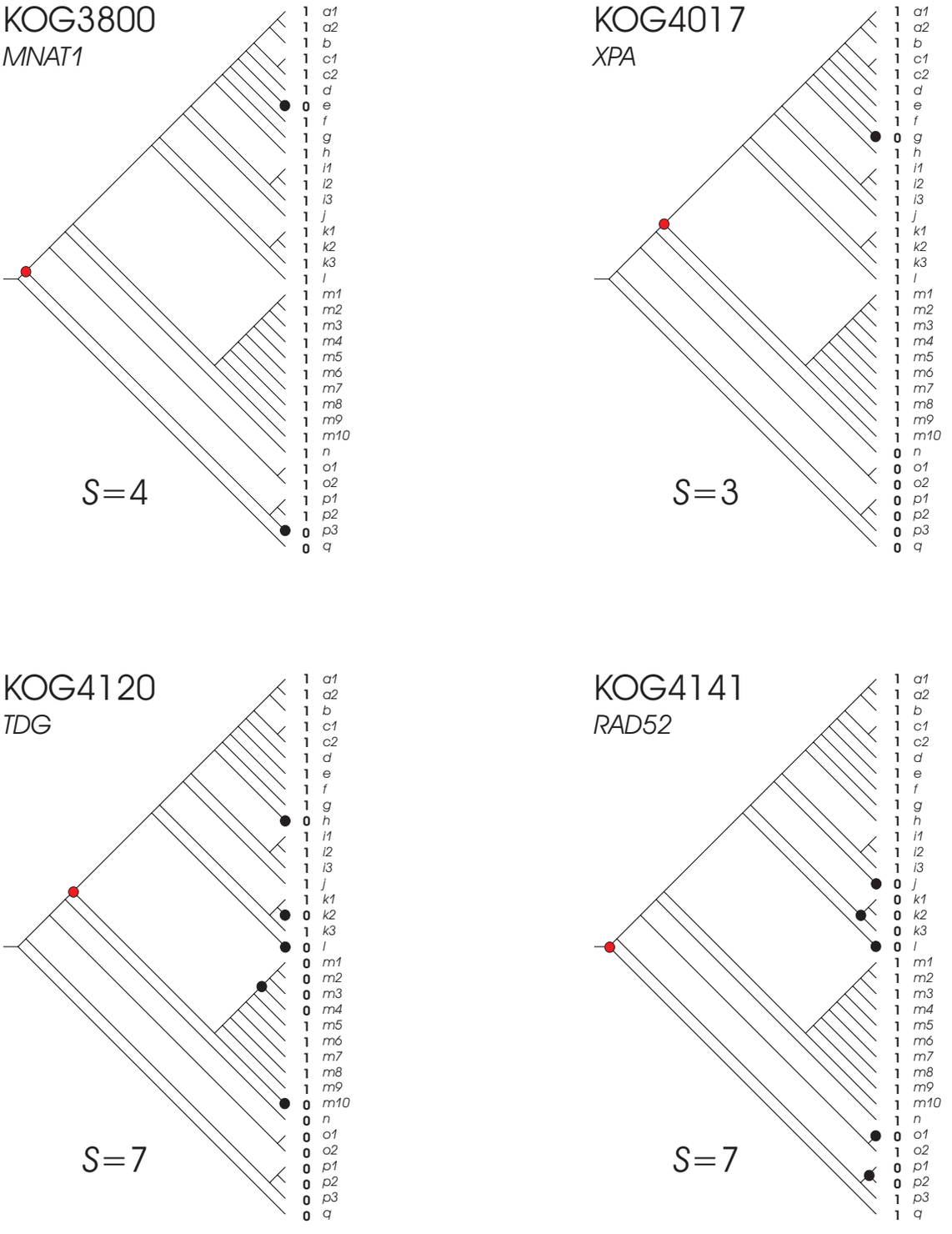
Supplementary Figure S34. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2807, KOG2841, KOG2875 and KOG2994, as in Supplementary Figure S14.



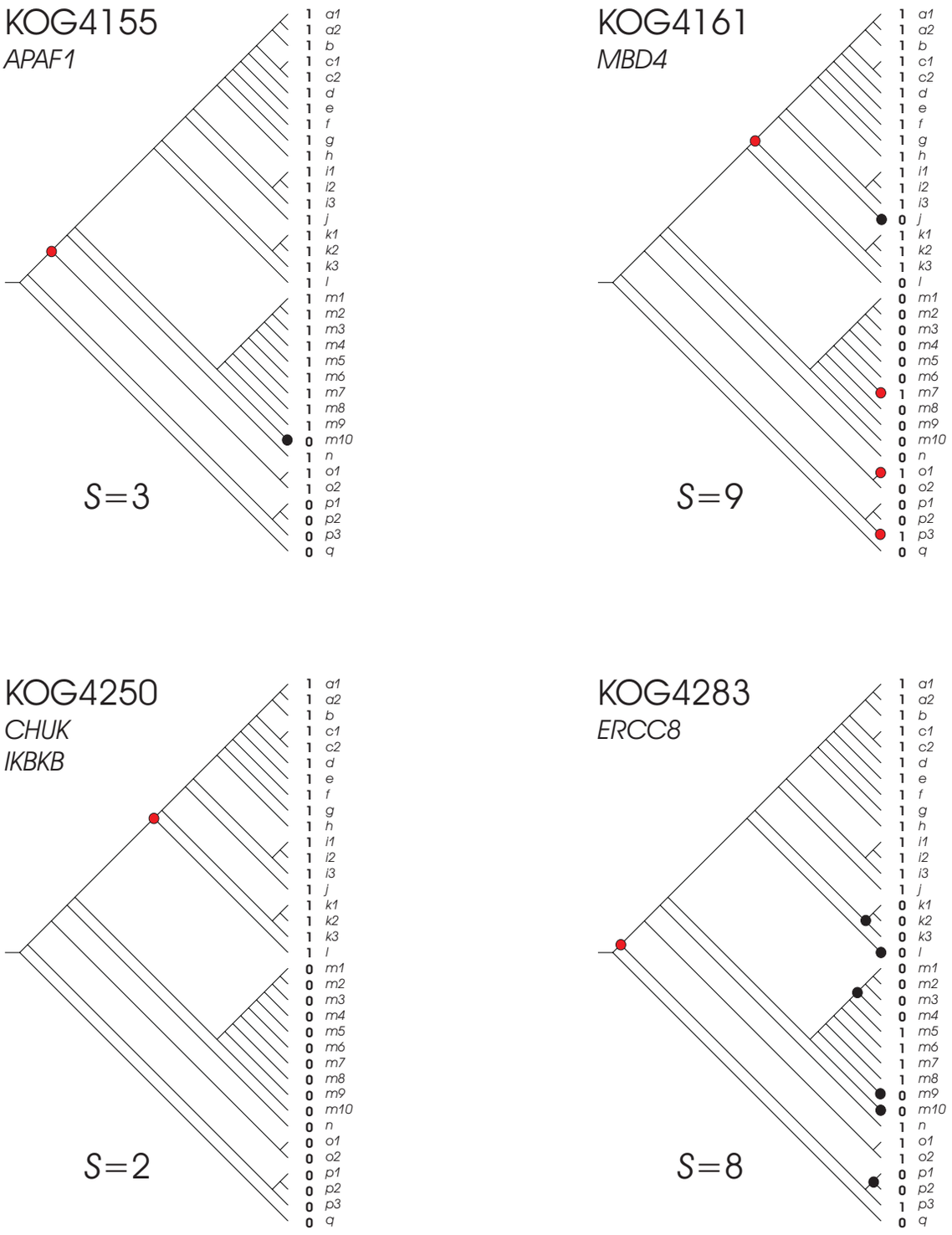
Supplementary Figure S35. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3108, KOG3194, KOG3226 and KOG3451, as in Supplementary Figure S14.



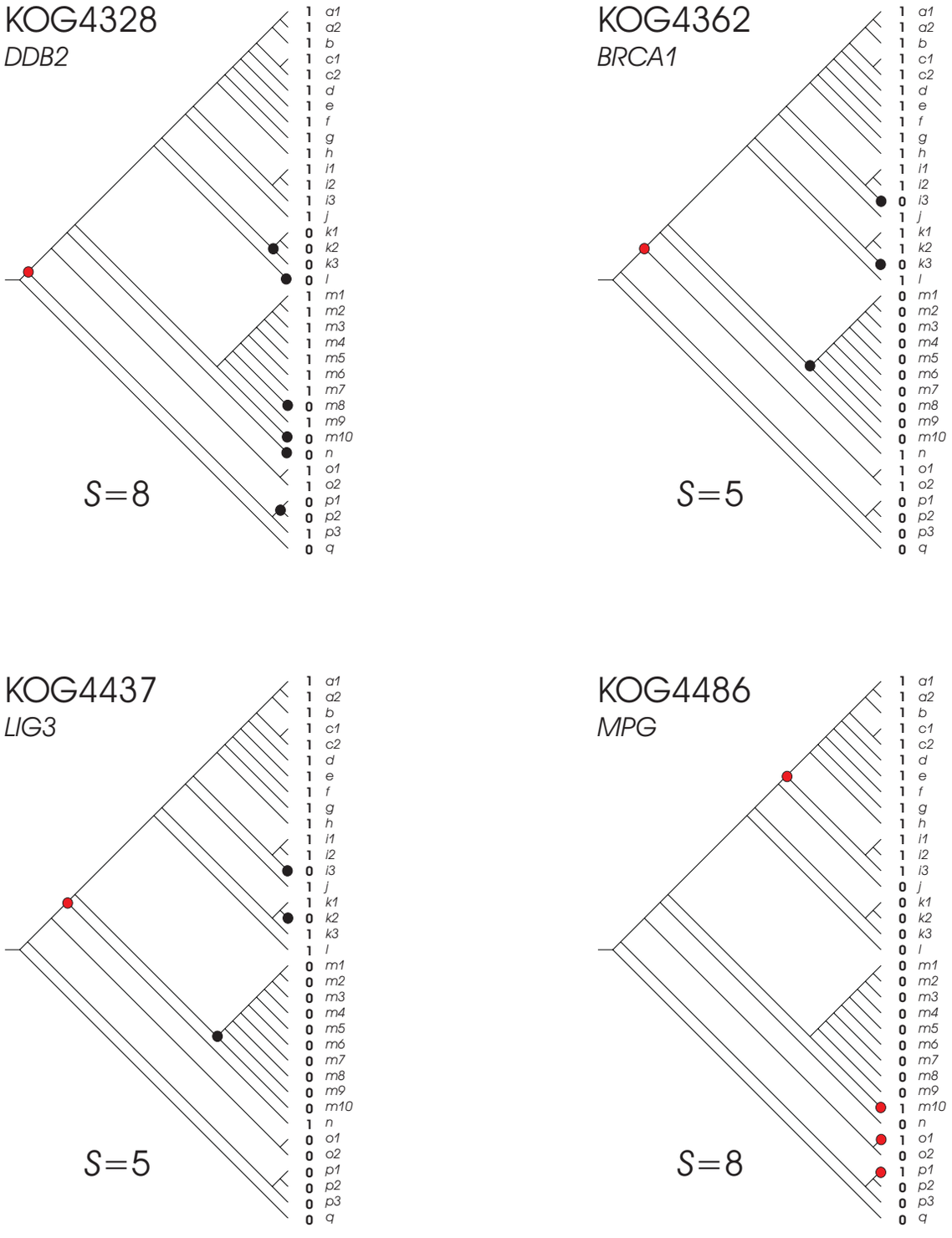
Supplementary Figure S36. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3453, KOG3471, KOG3573 and KOG3757, as in Supplementary Figure S14.



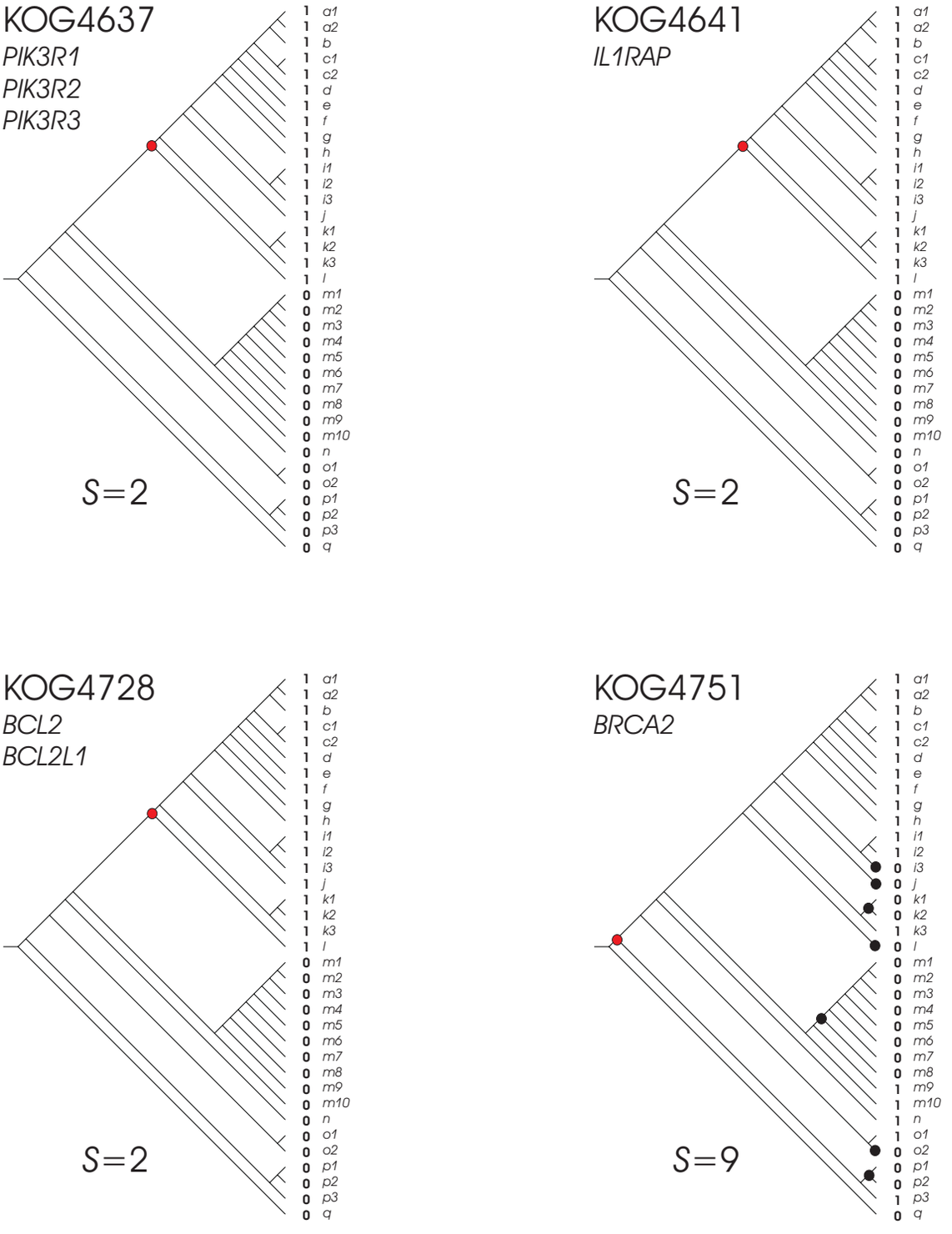
Supplementary Figure S37. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3800, KOG4017, KOG4120 and KOG4141, as in Supplementary Figure S14.



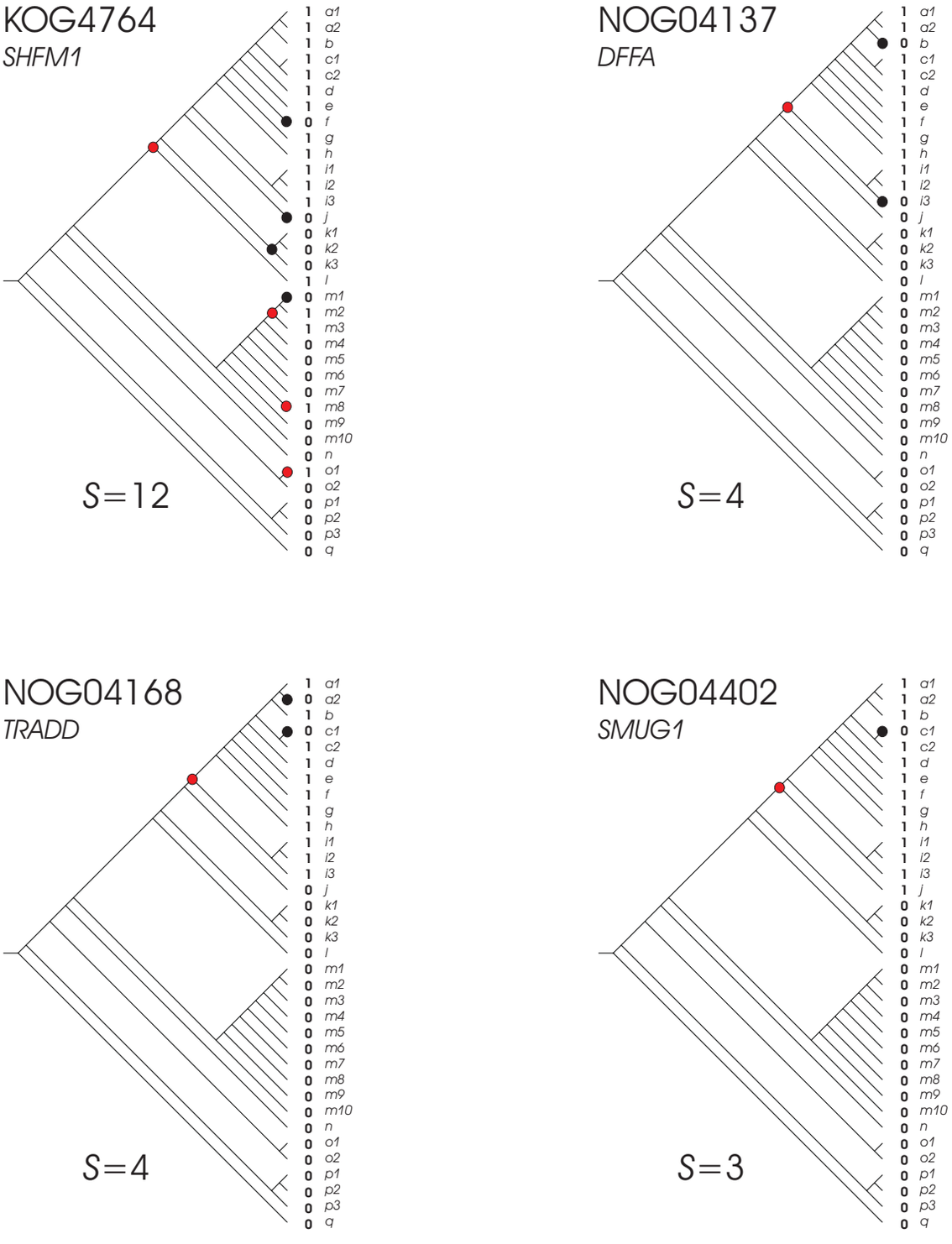
Supplementary Figure S38. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4155, KOG4161, KOG4250 and KOG4283, as in Supplementary Figure S14.



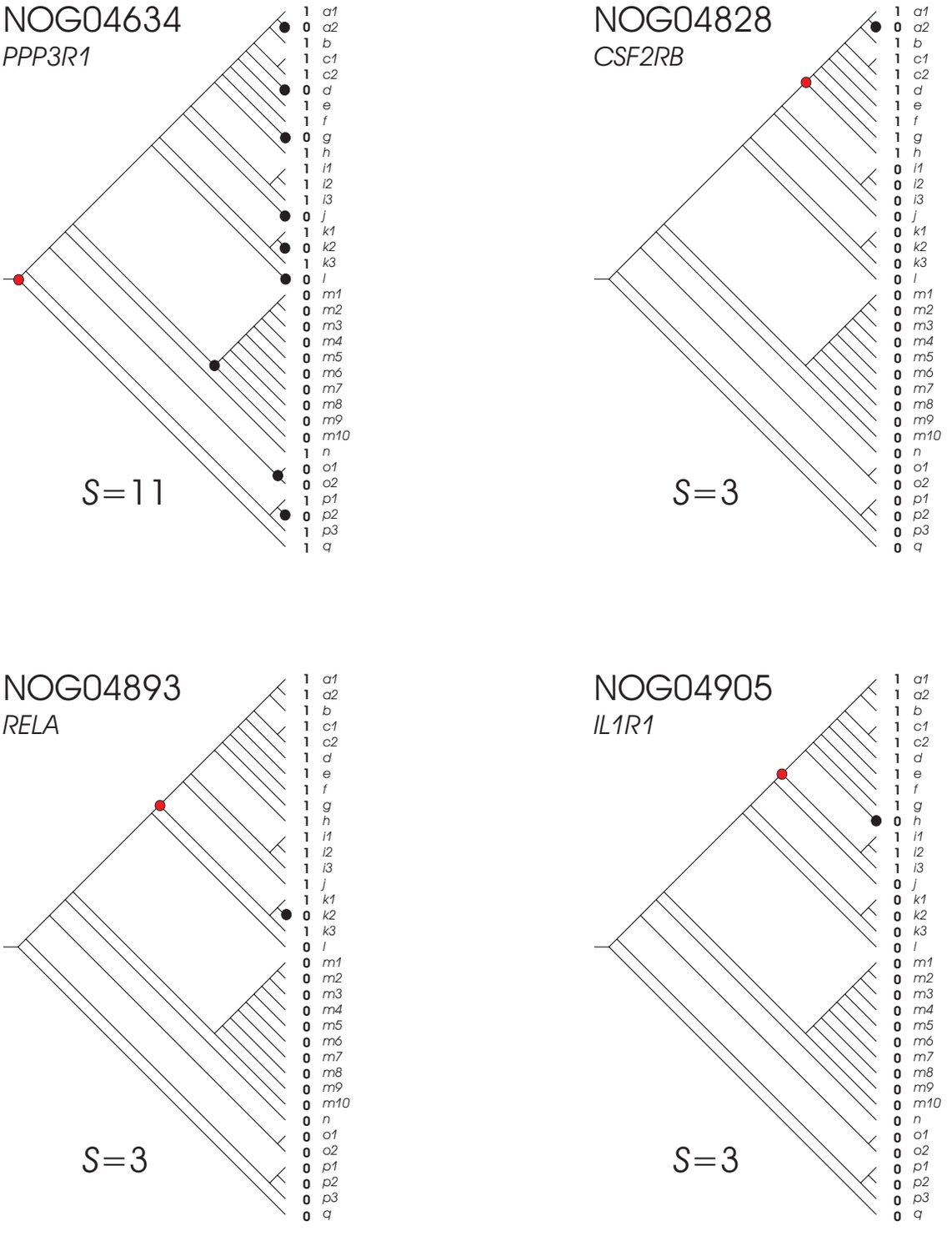
Supplementary Figure S39. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4328, KOG4362, KOG4437 and KOG4486, as in Supplementary Figure S14.



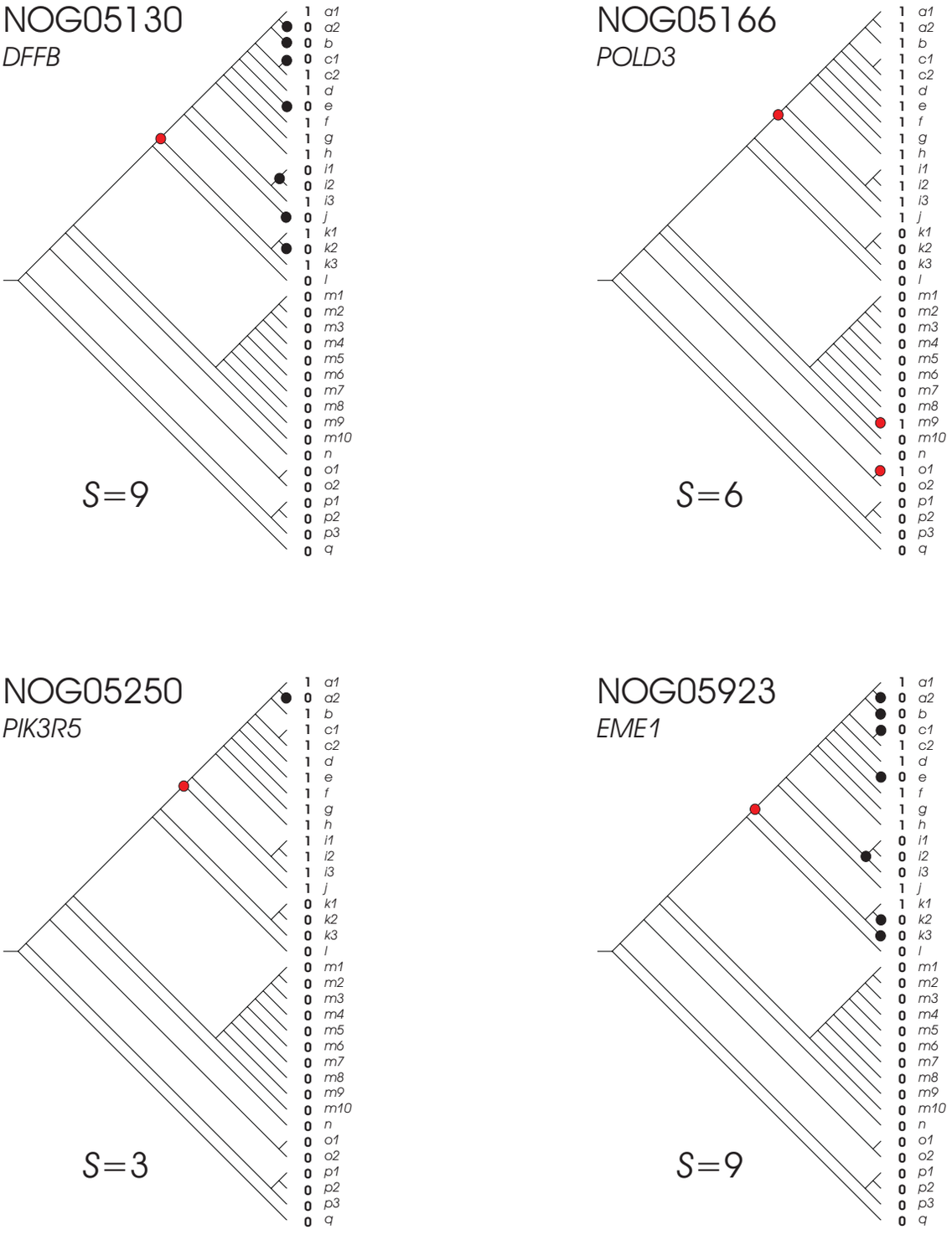
Supplementary Figure S40. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4637, KOG4641, KOG4728 and KOG4751, as in Supplementary Figure S14.



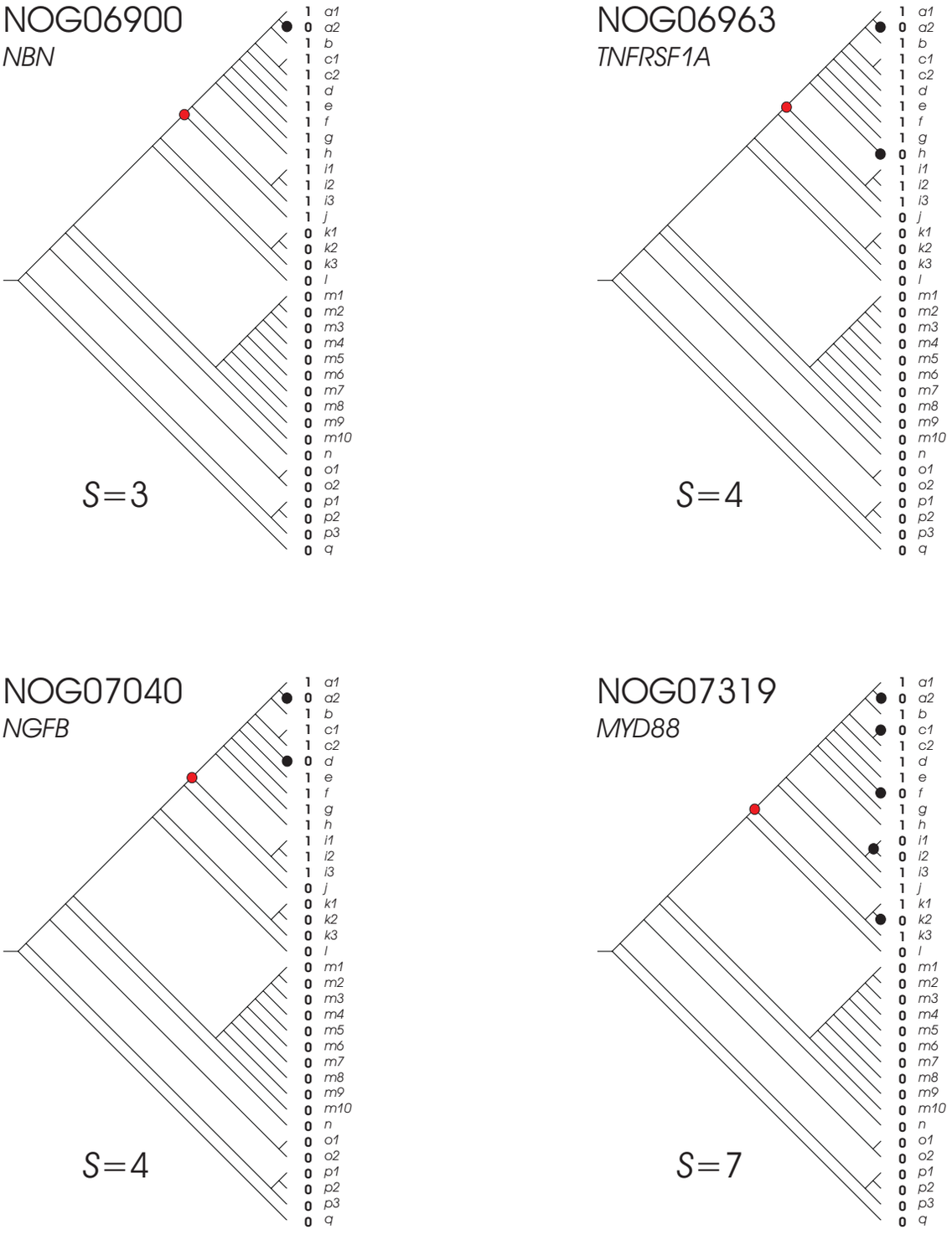
Supplementary Figure S41. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4764, NOG04137, NOG04168 and NOG04402, as in Supplementary Figure S14.



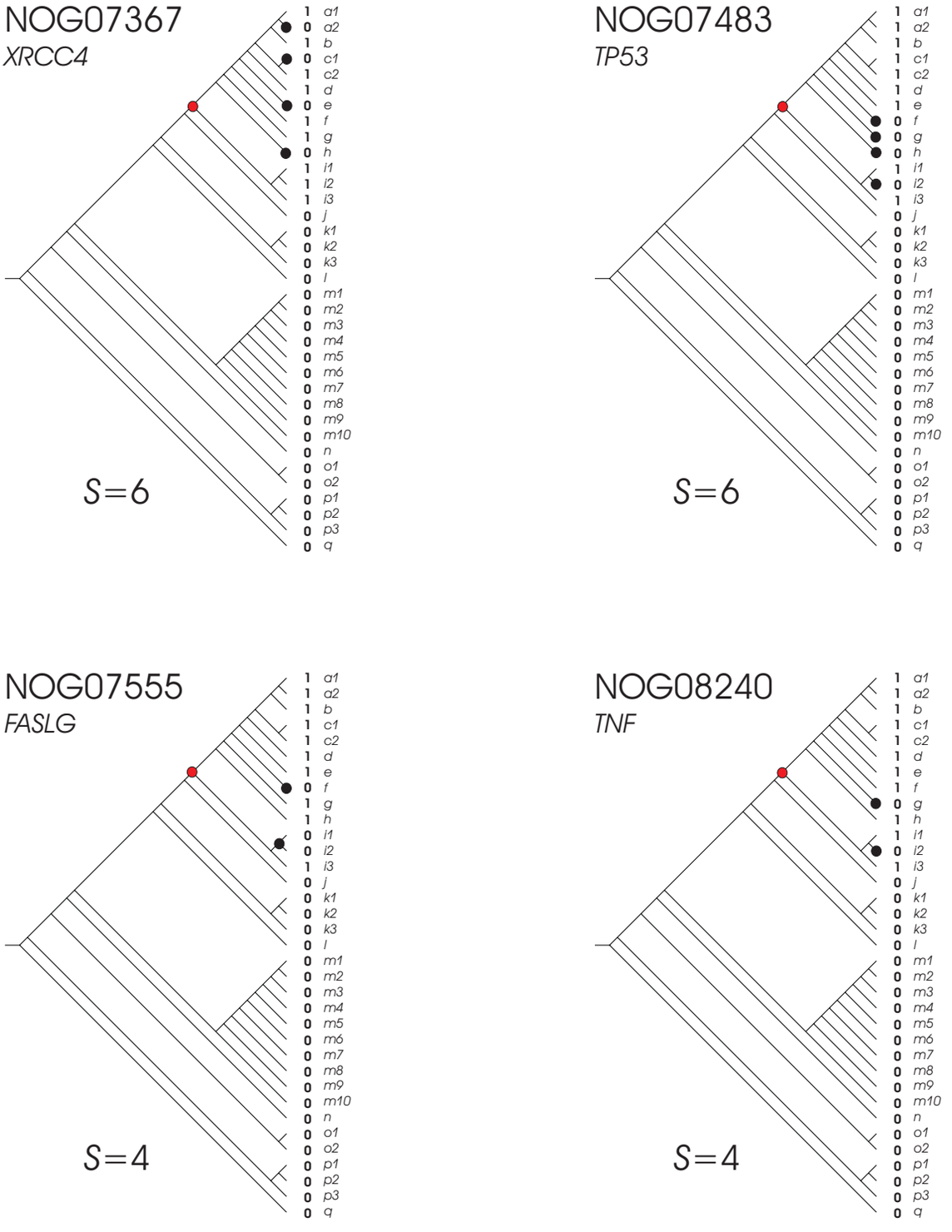
Supplementary Figure S42. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG04634, NOG04828, NOG04893 and NOG04905, as in Supplementary Figure S14.



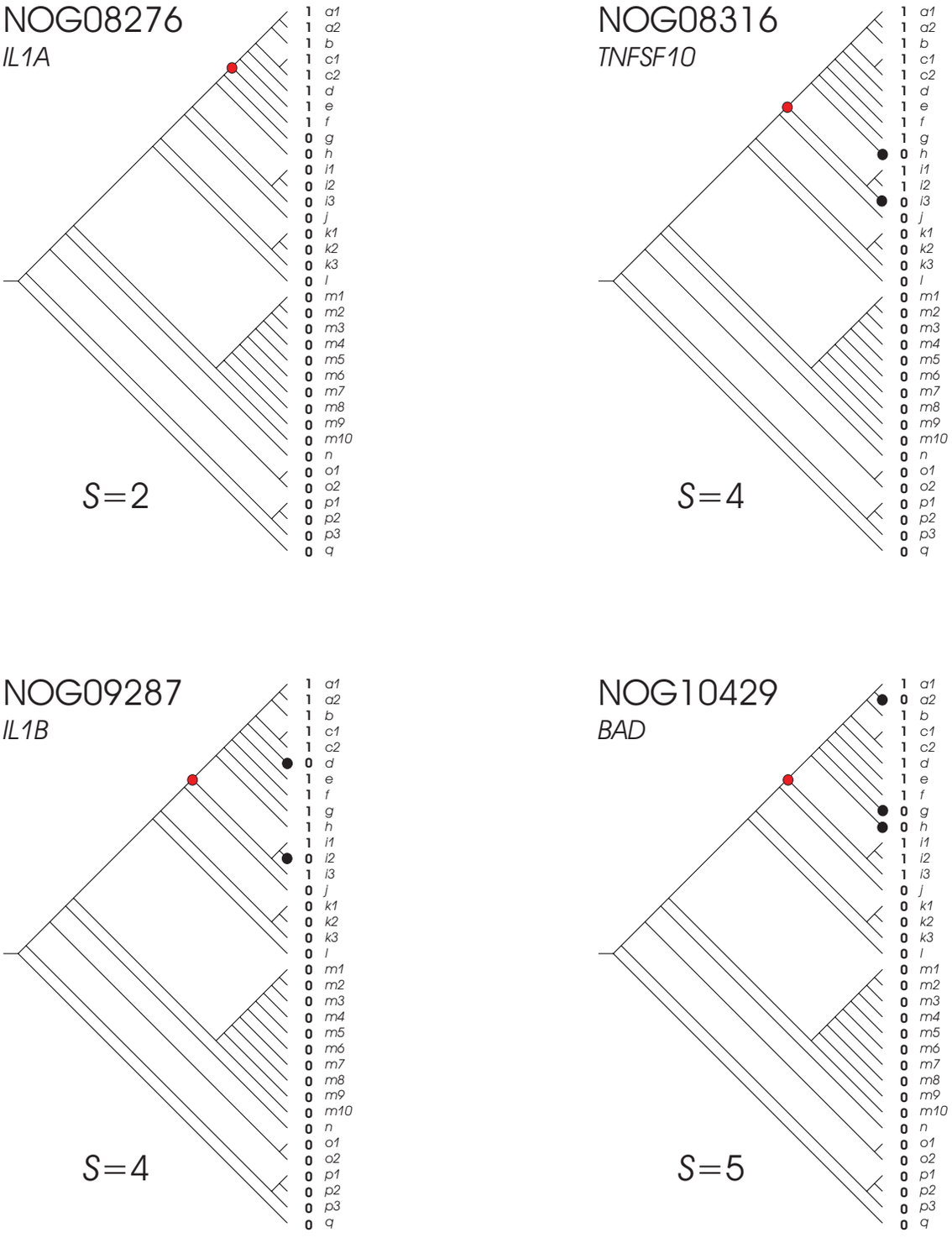
Supplementary Figure S43. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG05130, NOG05166, NOG05250 and NOG05923, as in Supplementary Figure S14.



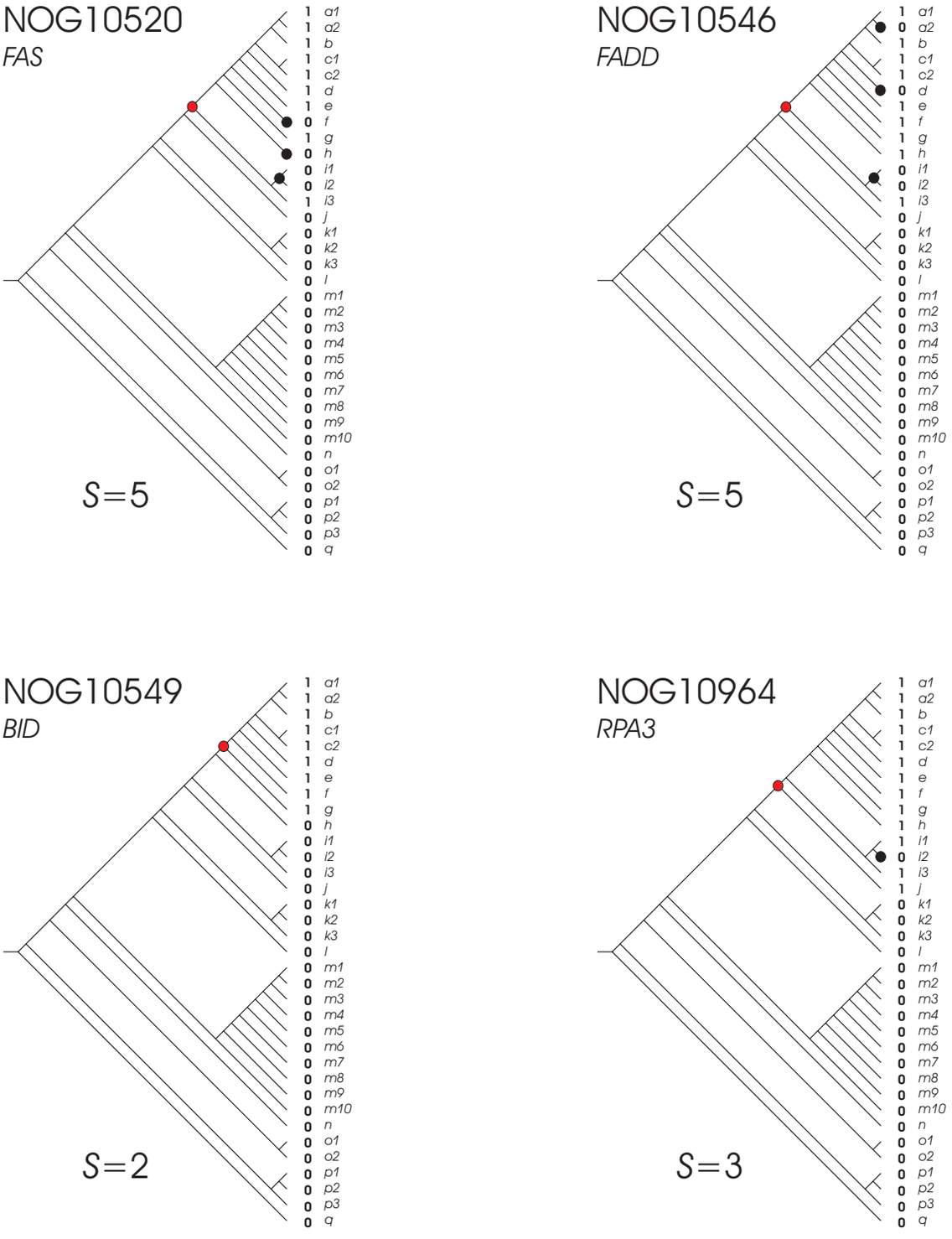
Supplementary Figure S44. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG06900, NOG06963, NOG07040 and NOG07319, as in Supplementary Figure S14.



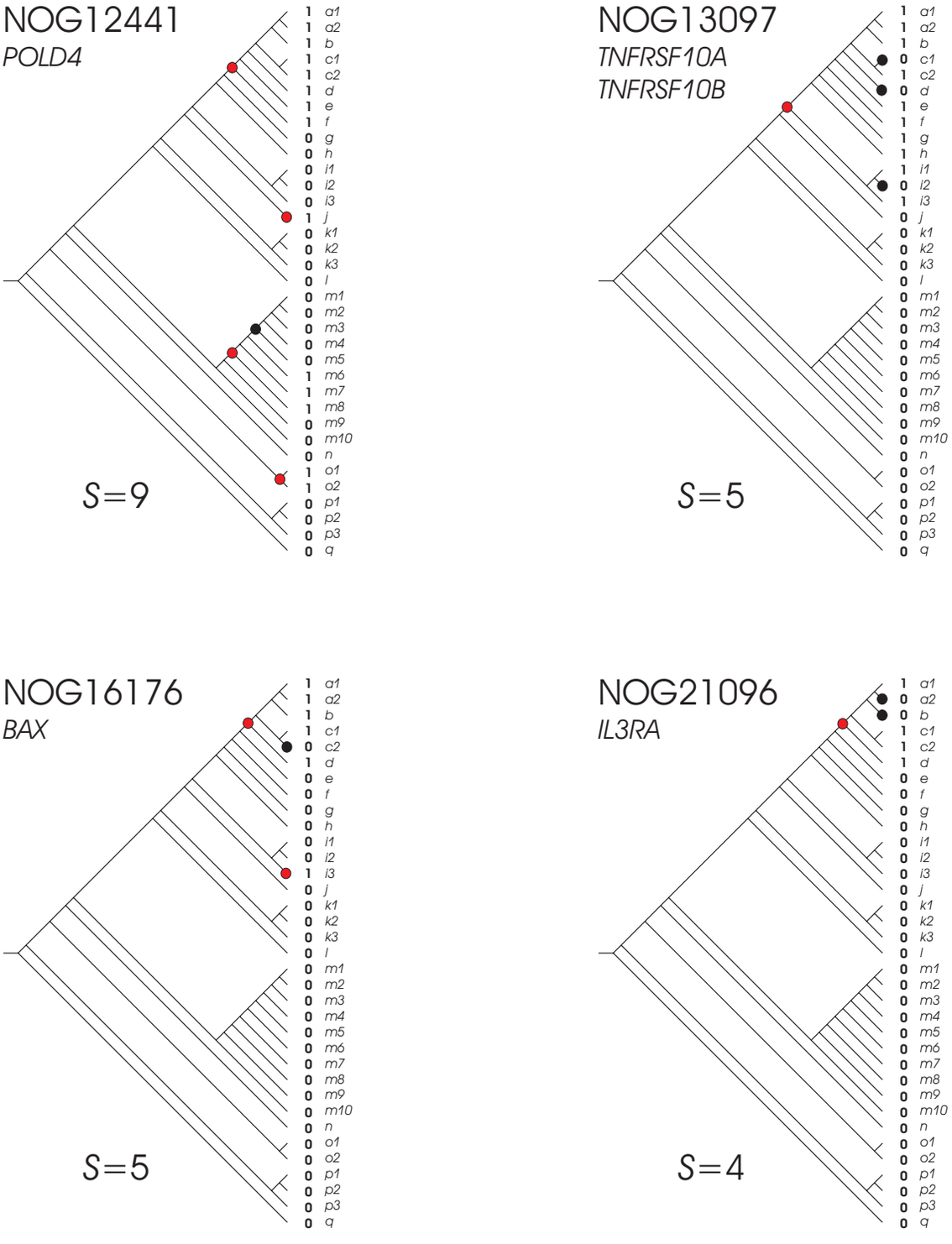
Supplementary Figure S45. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG07367, NOG07483, NOG07555 and NOG08240, as in Supplementary Figure S14.



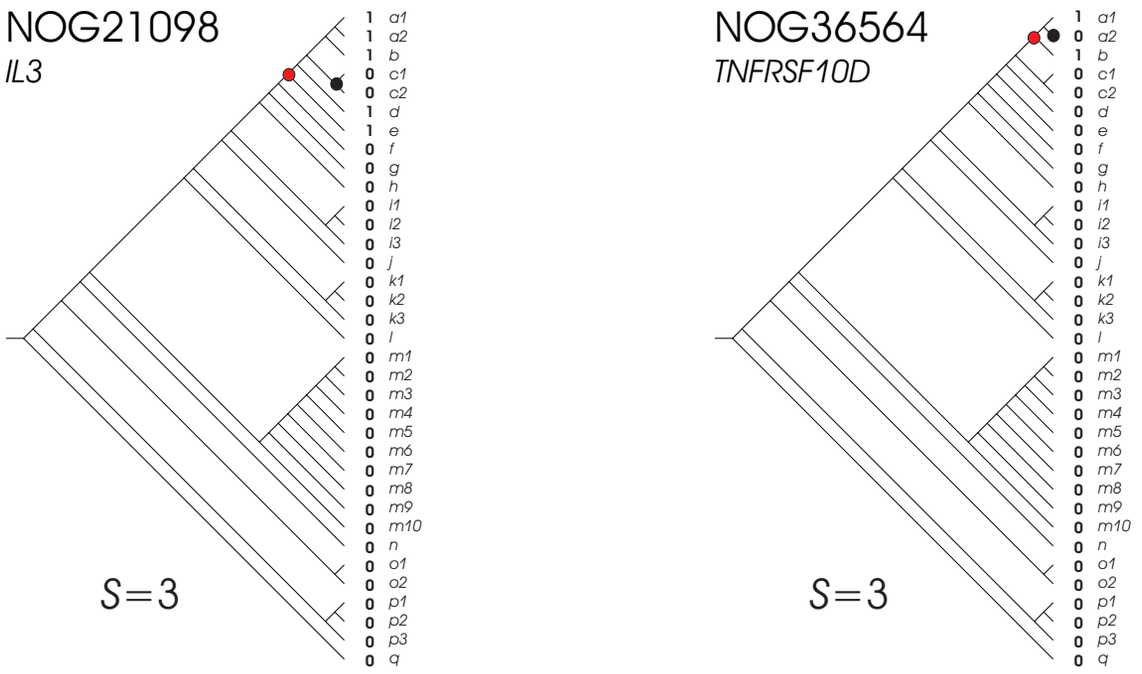
Supplementary Figure S46. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG08276, NOG08316, NOG09287 and NOG10429, as in Supplementary Figure S14.



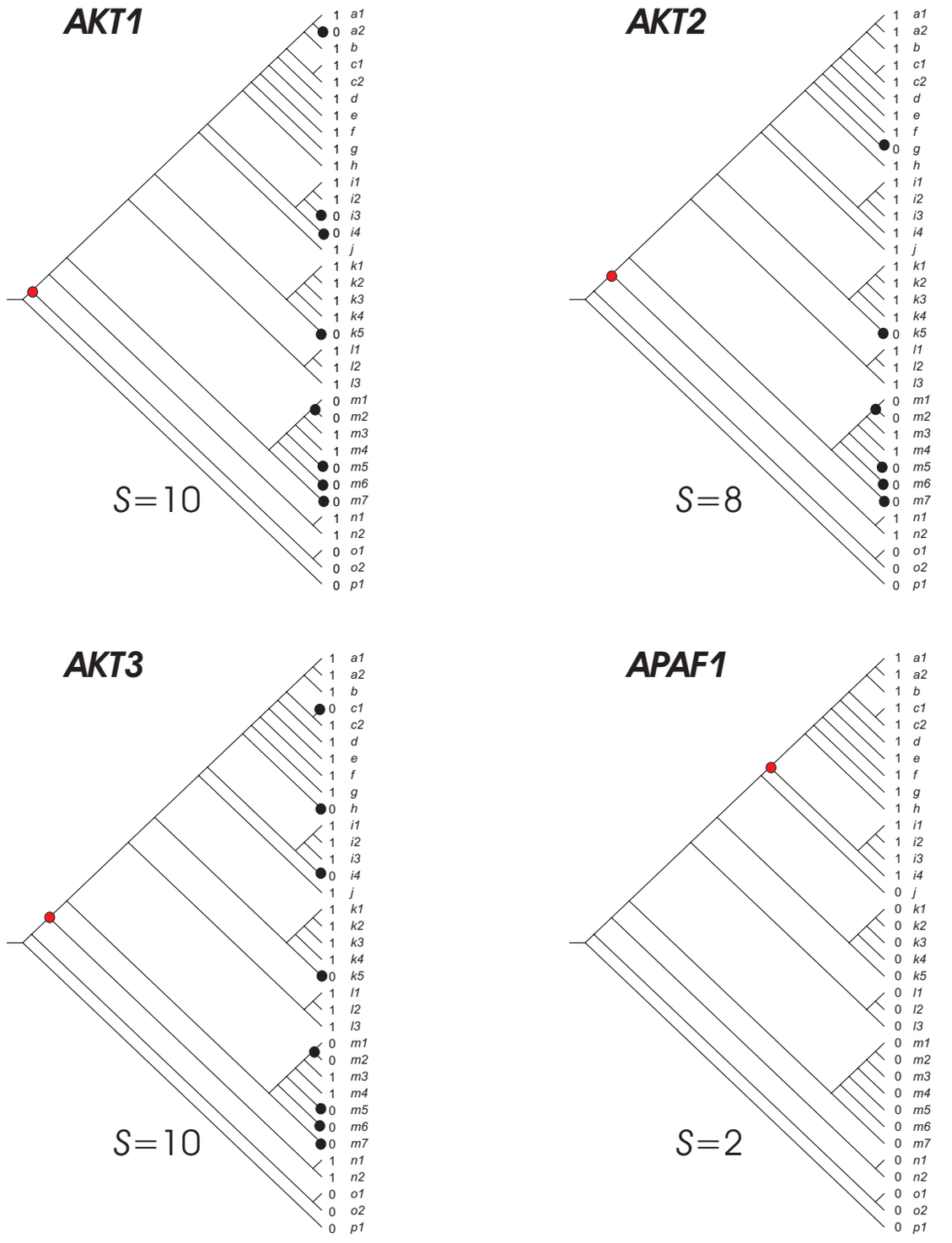
Supplementary Figure S47. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG10520, NOG10546, NOG10549 and NOG10964, as in Supplementary Figure S14.



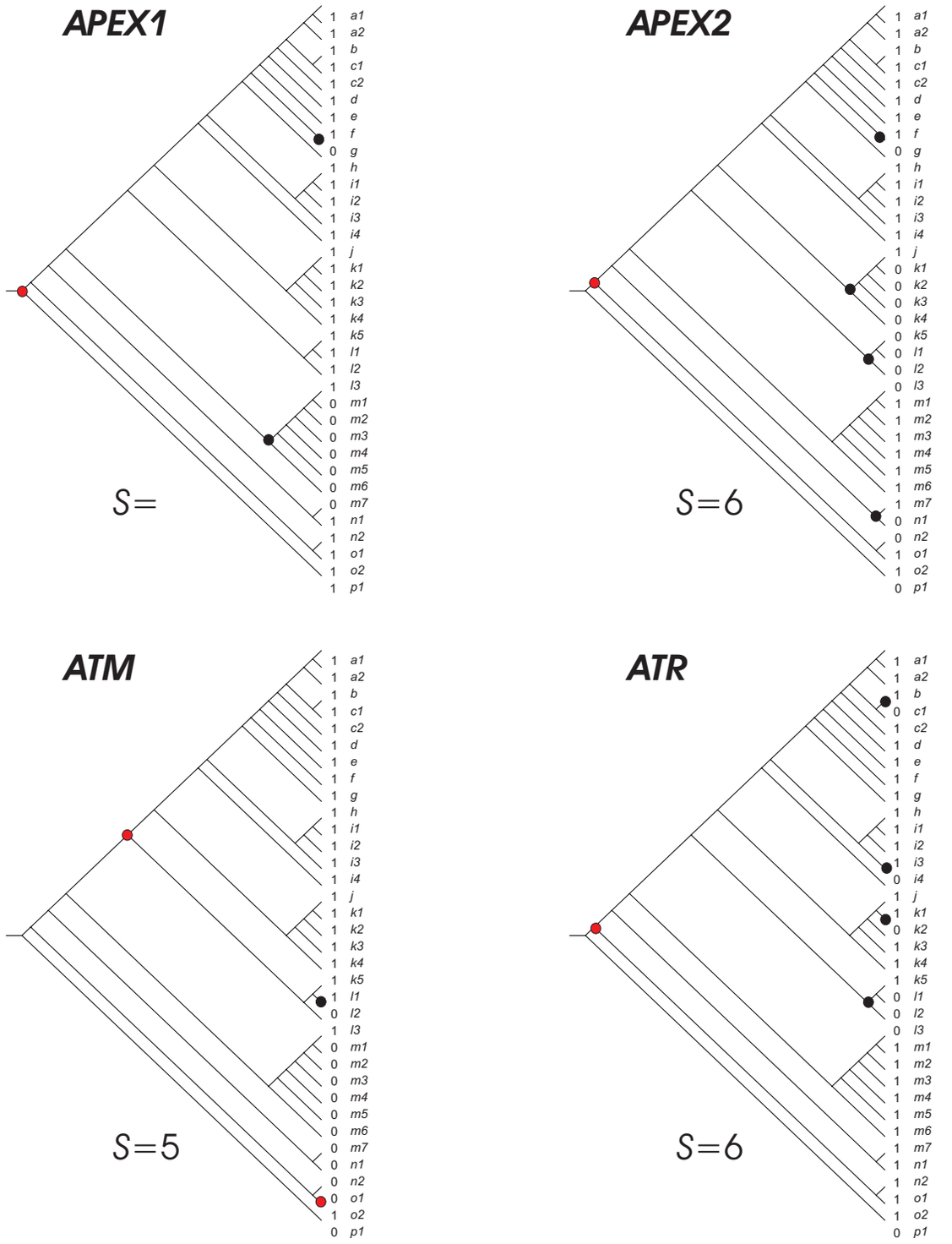
Supplementary Figure S48. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG12441, NOG13097, NOG16176 and NOG21096, as in Supplementary Figure S14.



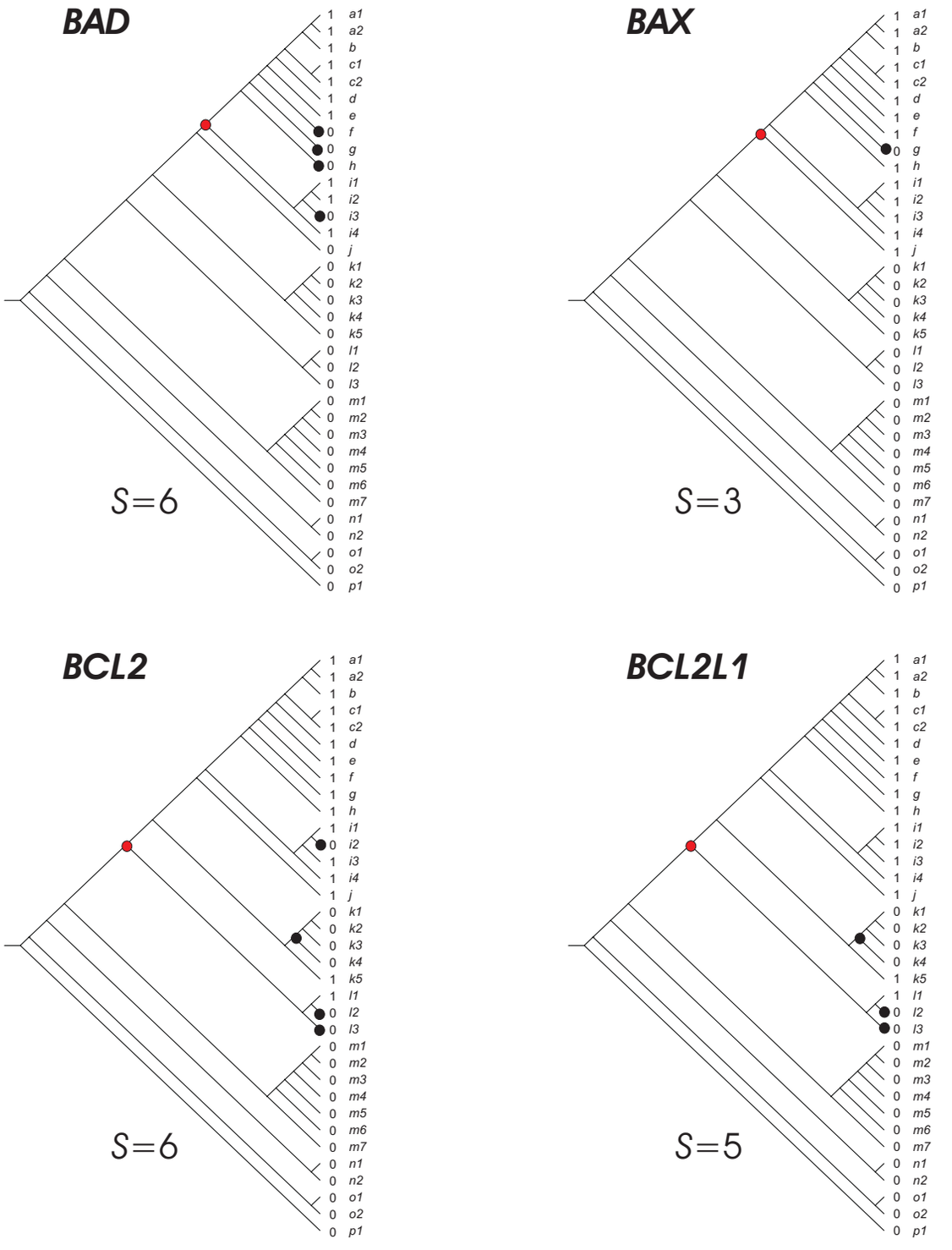
Supplementary Figure S49. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG21098 and NOG36564, as in Supplementary Figure S14.



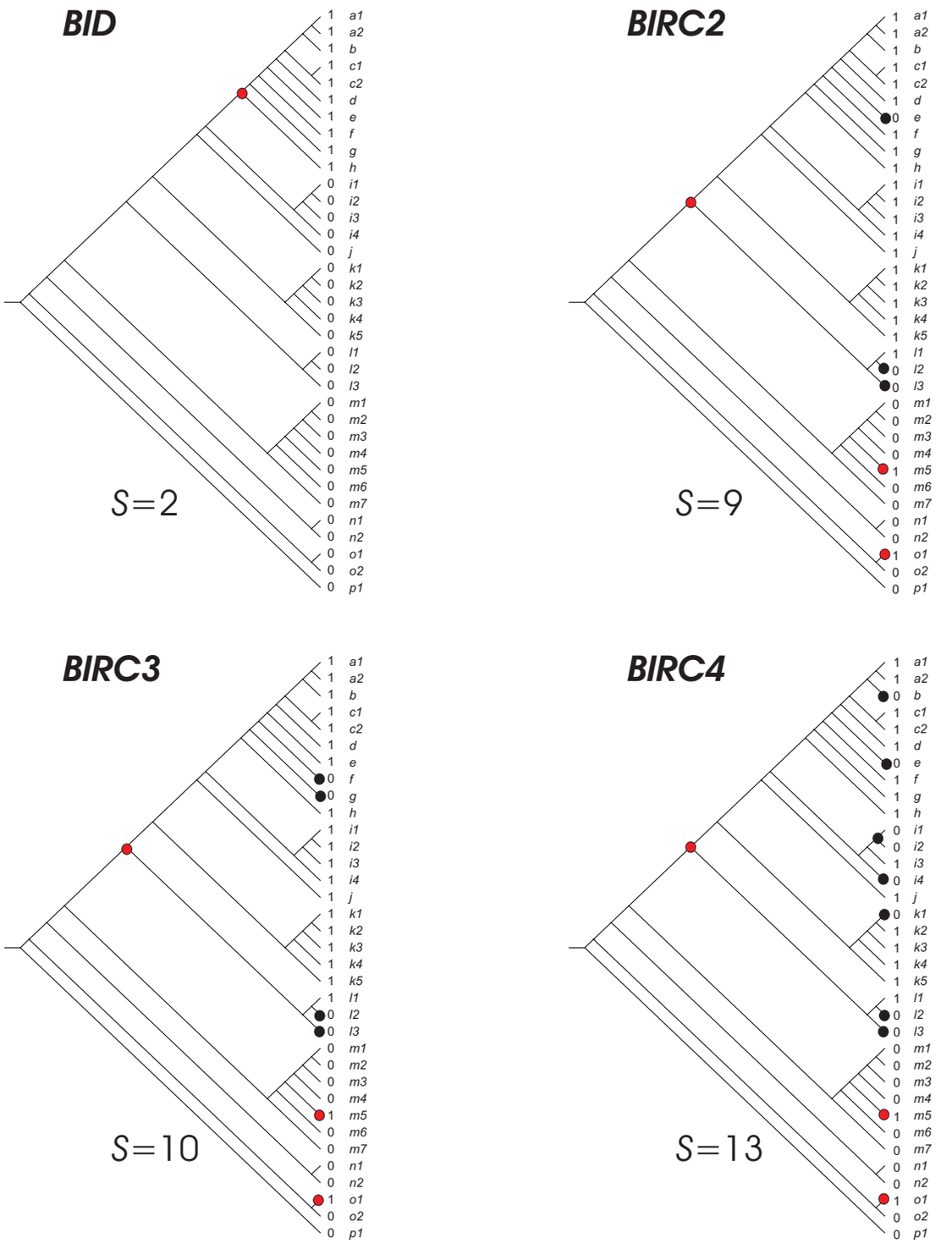
Supplementary Figure S50. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes: ***a1*** (*Homo sapiens*); ***a2*** (*Pan troglodytes*); ***b*** (*Macaca mulatta*); ***c1*** (*Rattus norvegicus*); ***c2*** (*Mus musculus*); ***d*** (*Canis familiaris*); ***e1*** (*Bos Taurus*); ***f*** (*Monodelphis domestica*); ***g*** (*Gallus gallus*); ***h*** (*Xenopus tropicalis*); ***i1*** (*Takifugu rubripes*); ***i2*** (*Tetraodon nigroviridis*); ***i3*** (*Gasterosteus aculeatus*); ***i4*** (*Danio rerio*); ***j*** (*Ciona intestinalis*); ***k1*** (*Drosophila melanogaster*); ***k2*** (*Drosophila pseudoobscura*); ***k3*** (*Anopheles gambiae*); ***k4*** (*Aedes aegypti*); ***k5*** (*Apis mellifera*); ***l1*** (*Caenorhabditis briggsae*); ***l2*** (*Caenorhabditis remanei*); ***l3*** (*Caenorhabditis elegans*); ***m1*** (*Kluyveromyces lactis*); ***m2*** (*Saccharomyces cerevisiae*); ***m3*** (*Candida glabrata*); ***m4*** (*Debaryomyces hansenii*); ***m5*** (*Yarrowia lipolytica*); ***m6*** (*Schizosaccharomyces pombe*); ***m7*** (*Cryptococcus neoformans*); ***n1*** (*Entamoeba histolytica*); ***n2*** (*Dictyostelium discoideum*); ***o1*** (*Arabidopsis thaliana*); ***o2*** (*Oryza sativa*); ***p1*** (*Escherichia coliK12*);



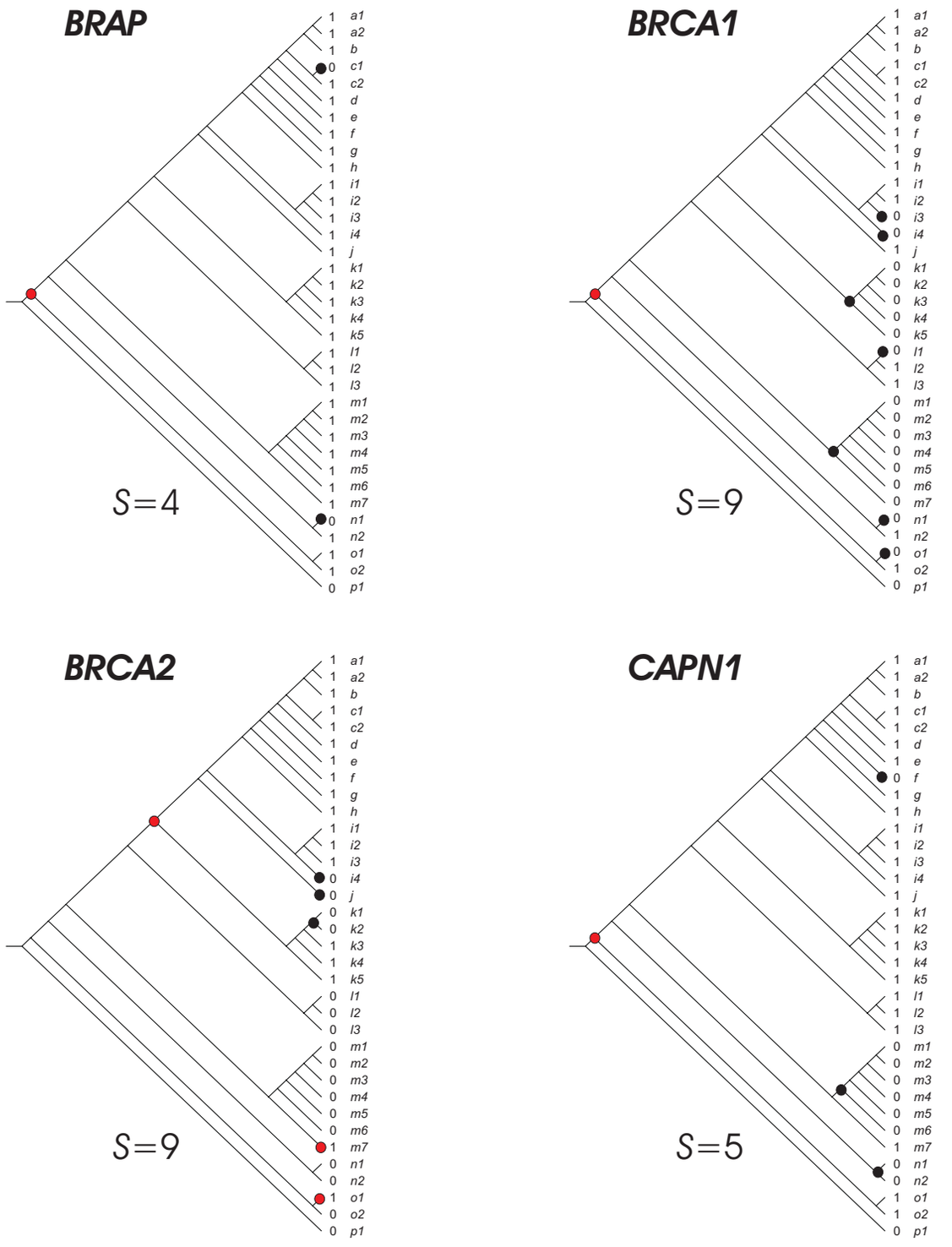
Supplementary Figure S51. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



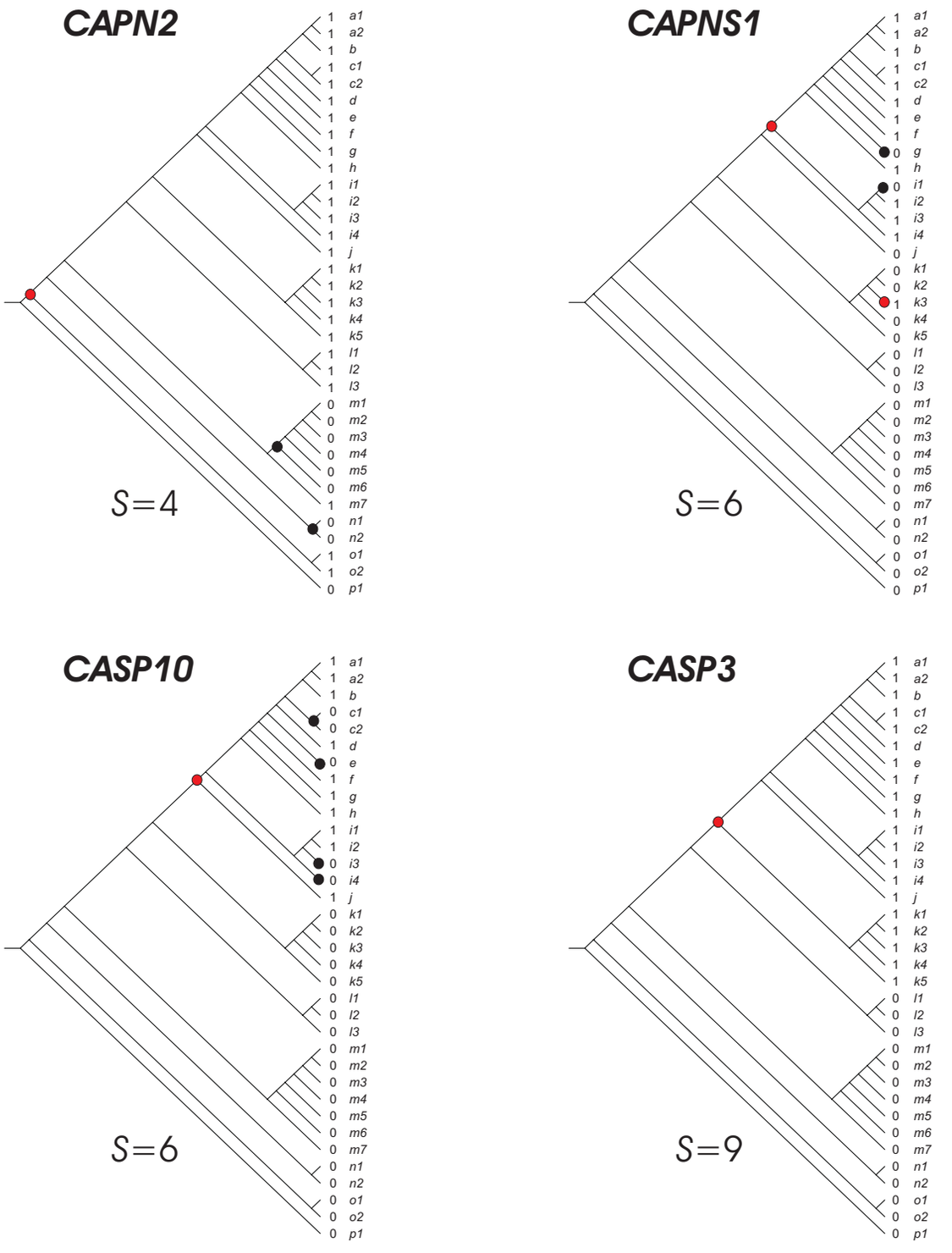
Supplementary Figure S52. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



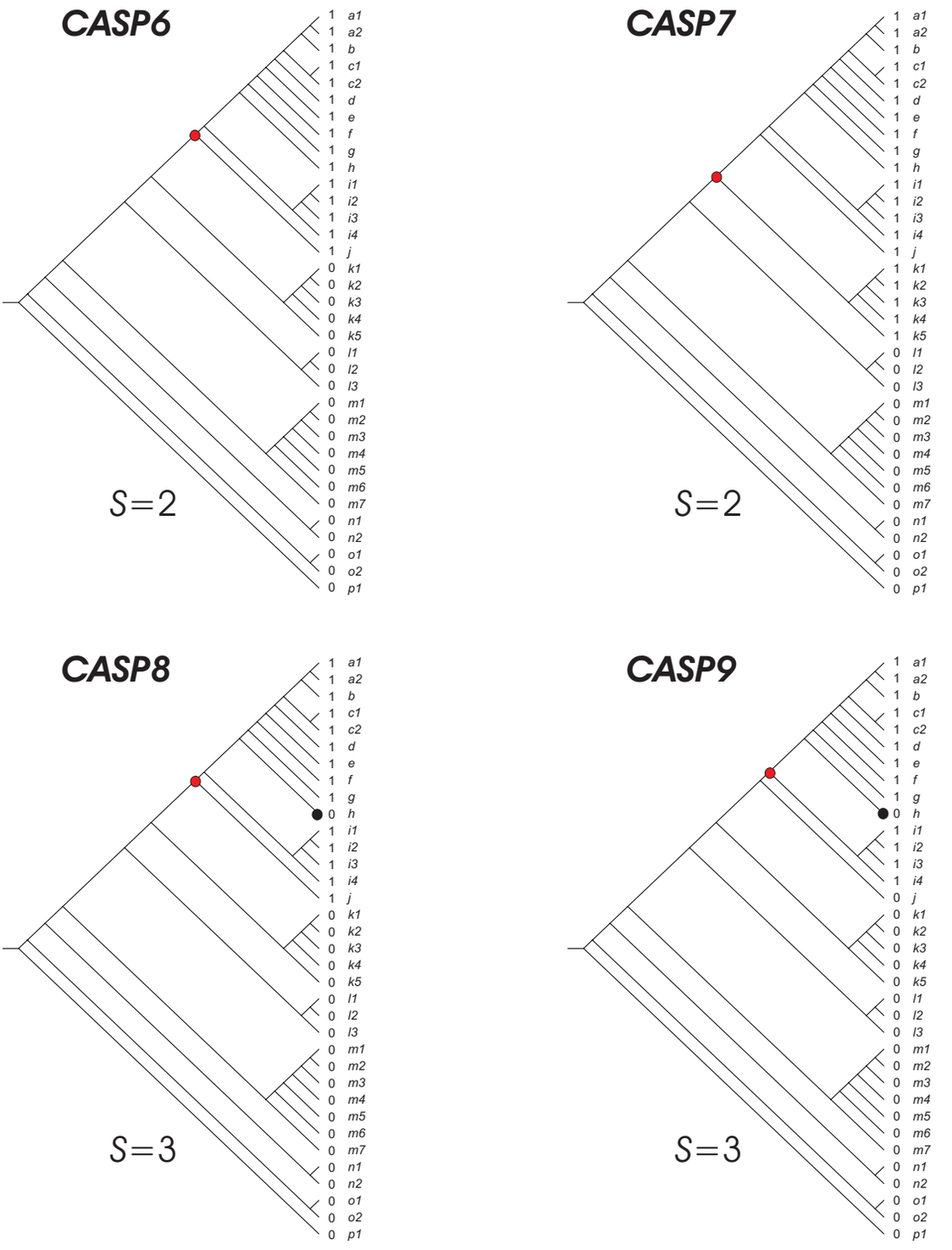
Supplementary Figure S53. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



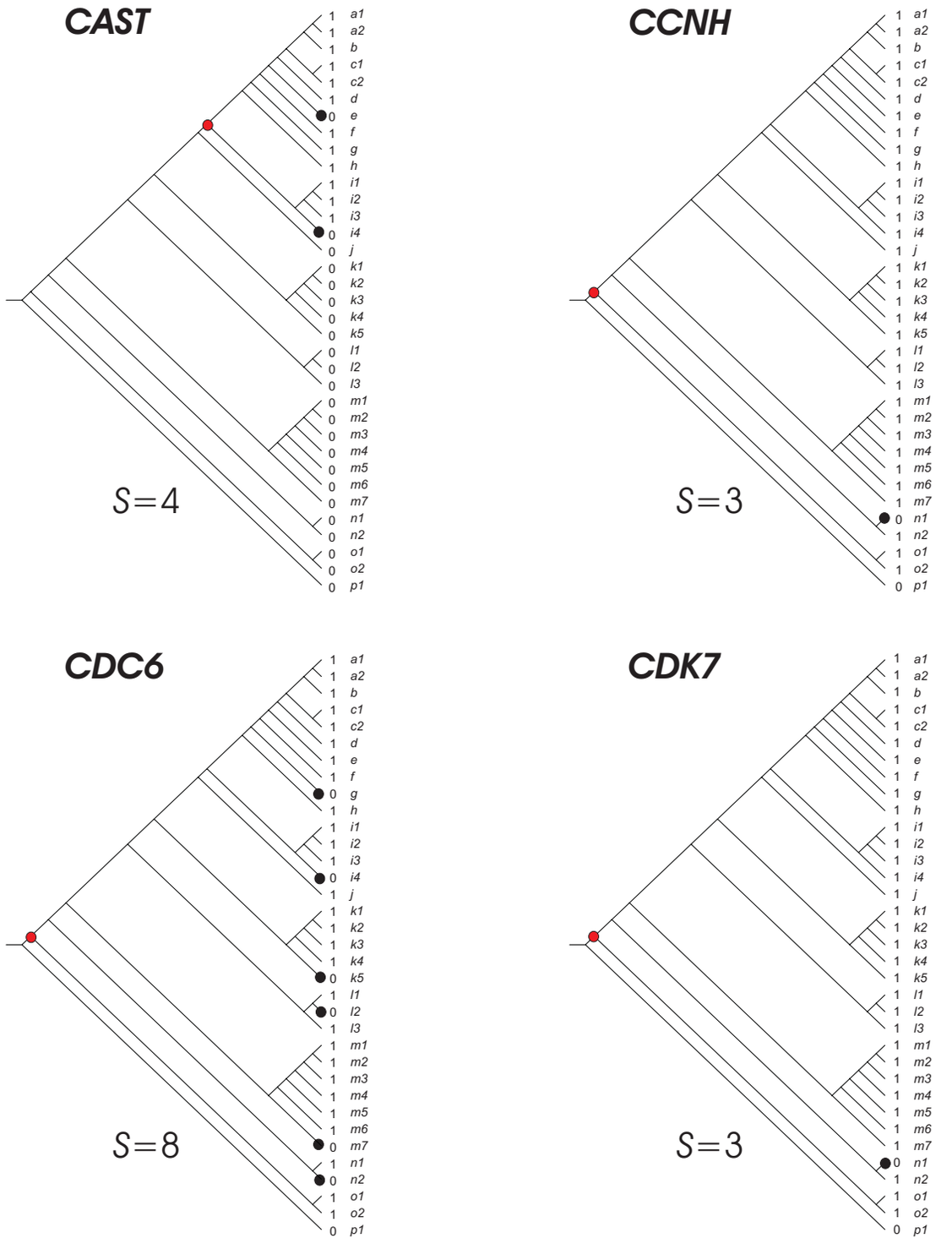
Supplementary Figure S54. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



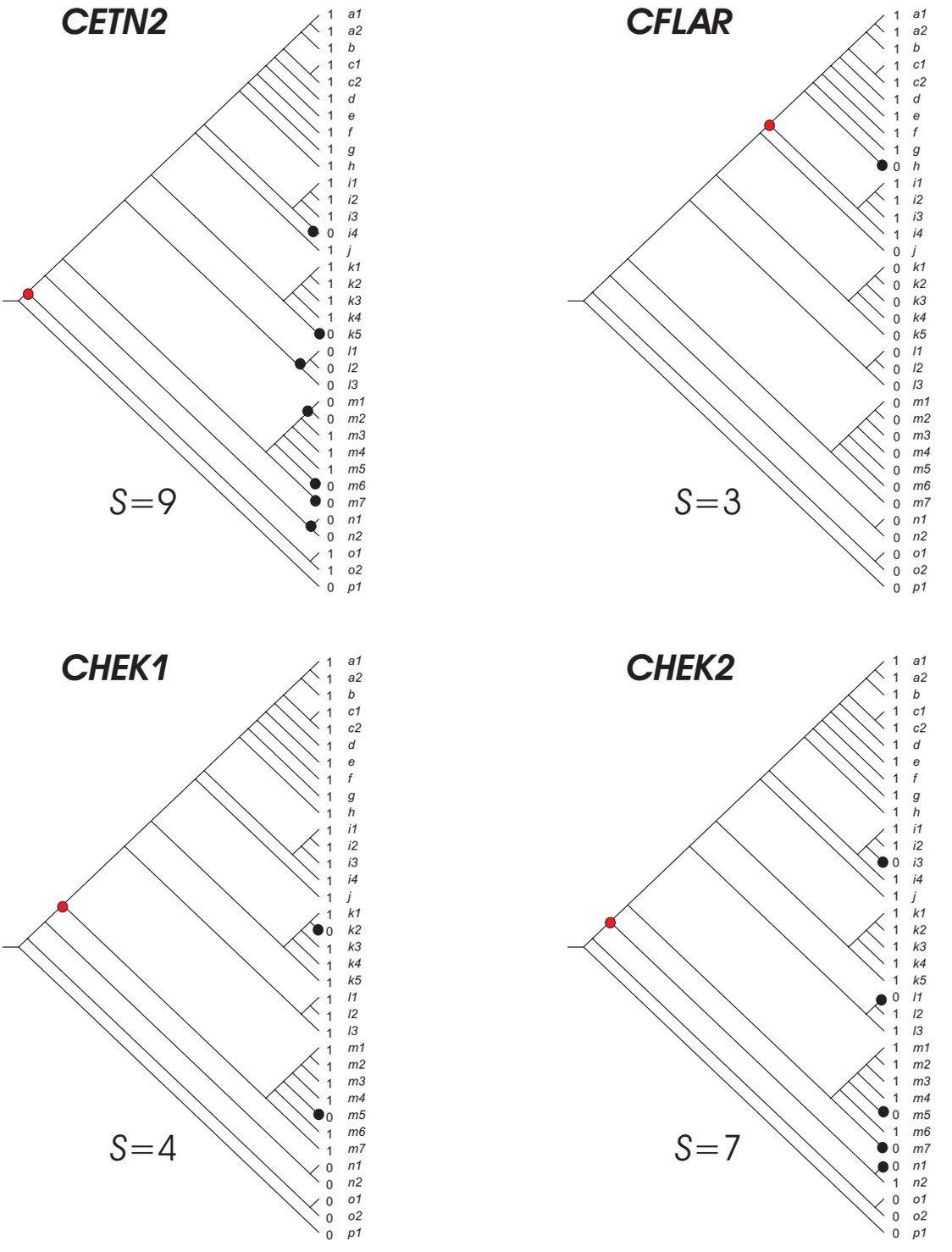
Supplementary Figure S55. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



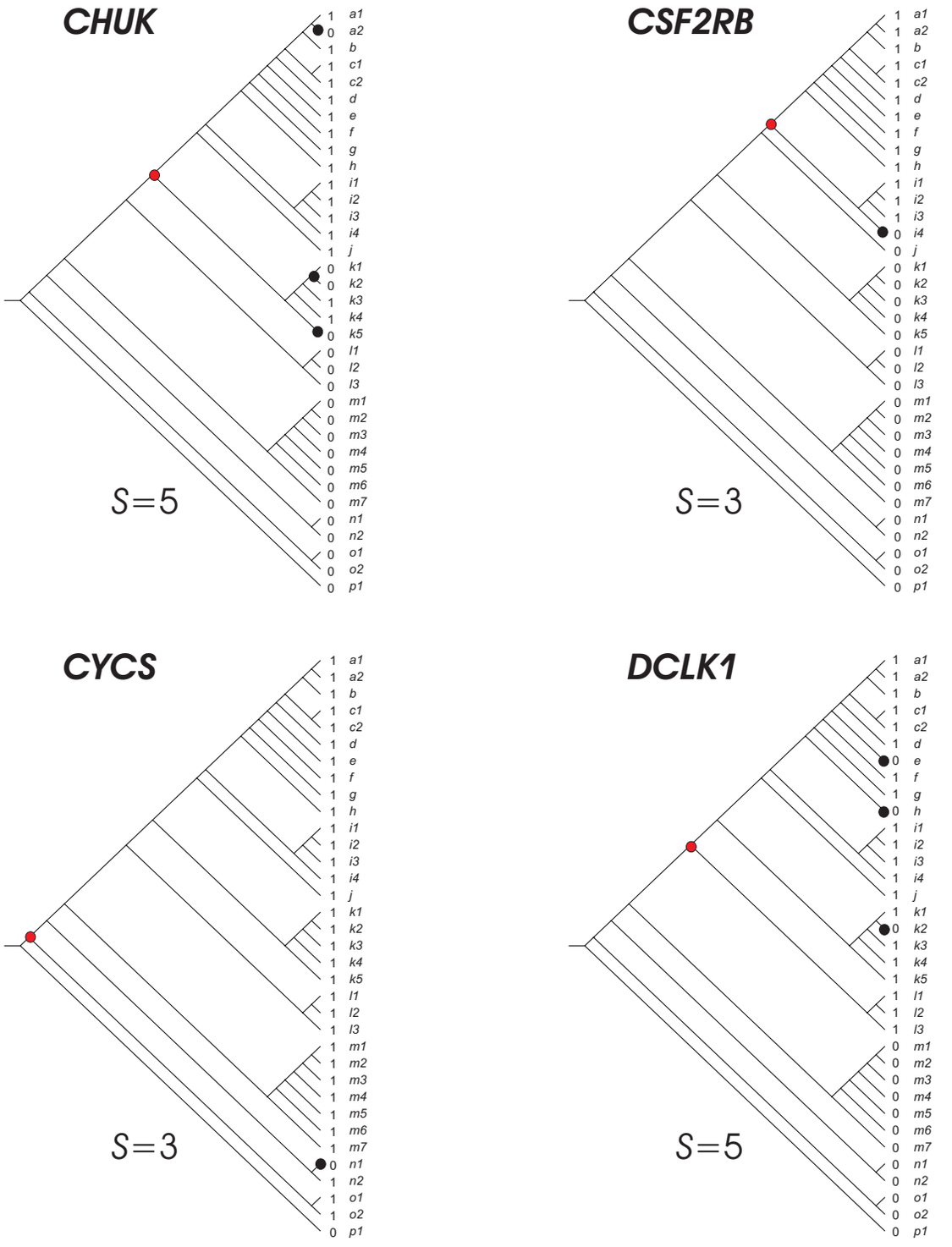
Supplementary Figure S56. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



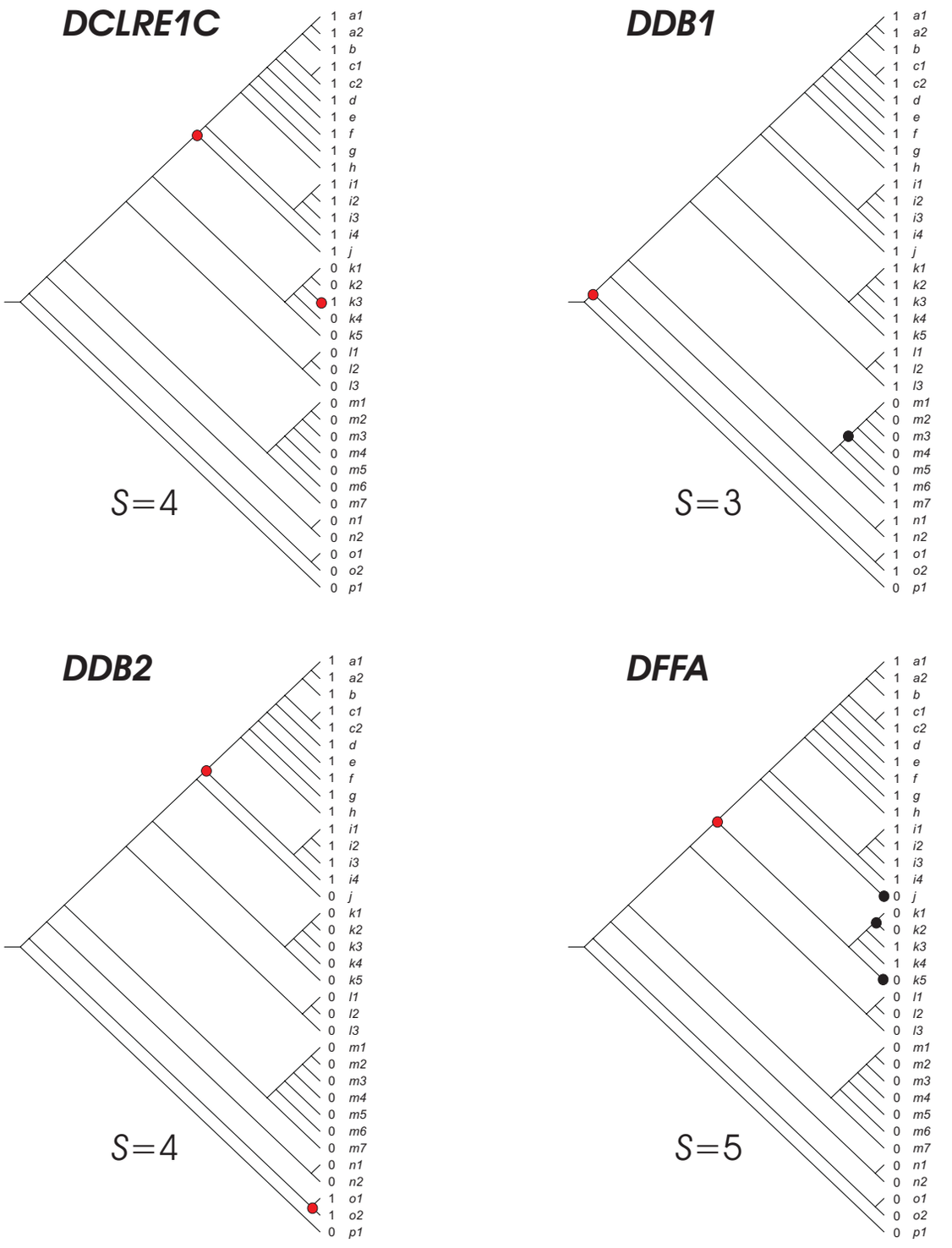
Supplementary Figure S57. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



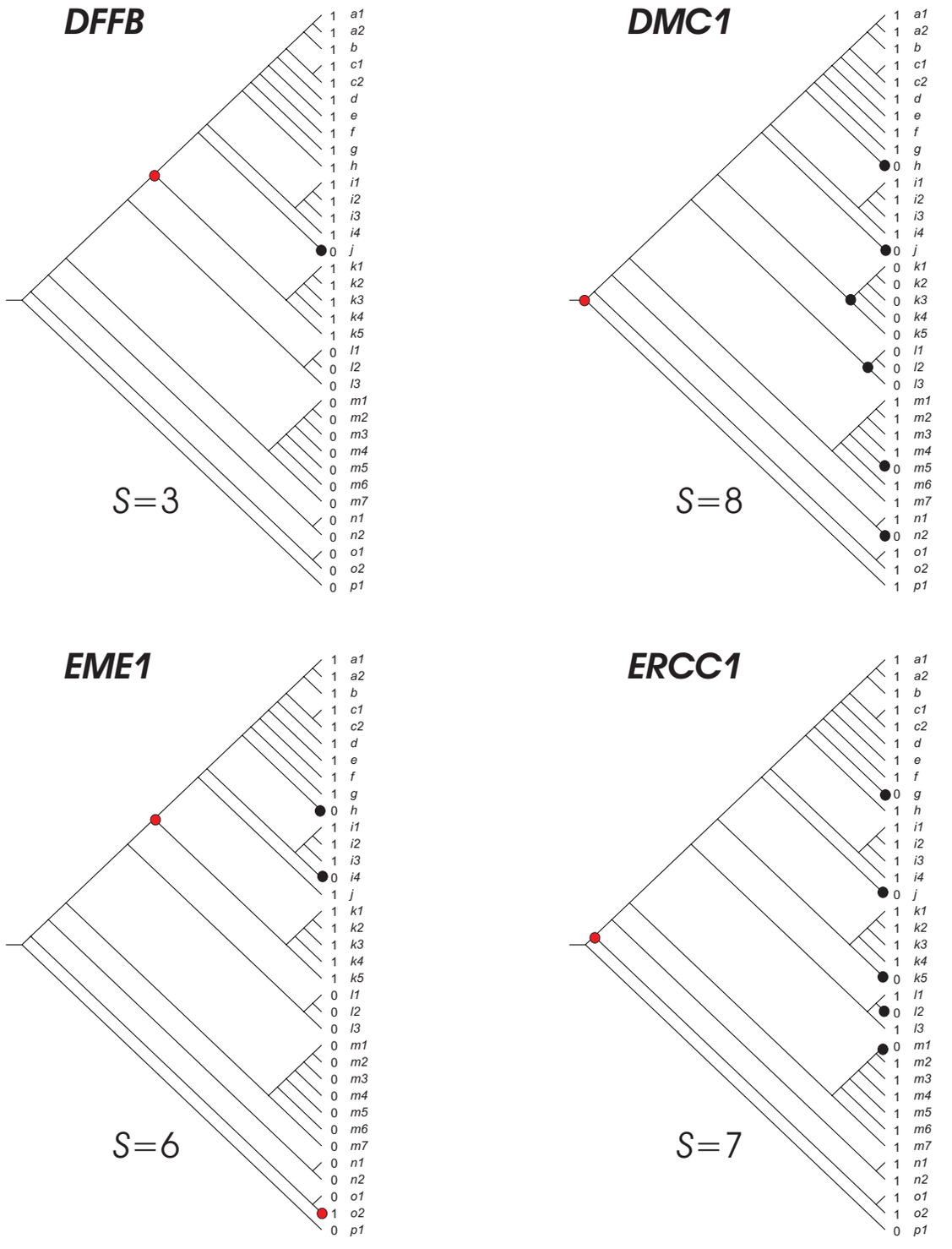
Supplementary Figure S58. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



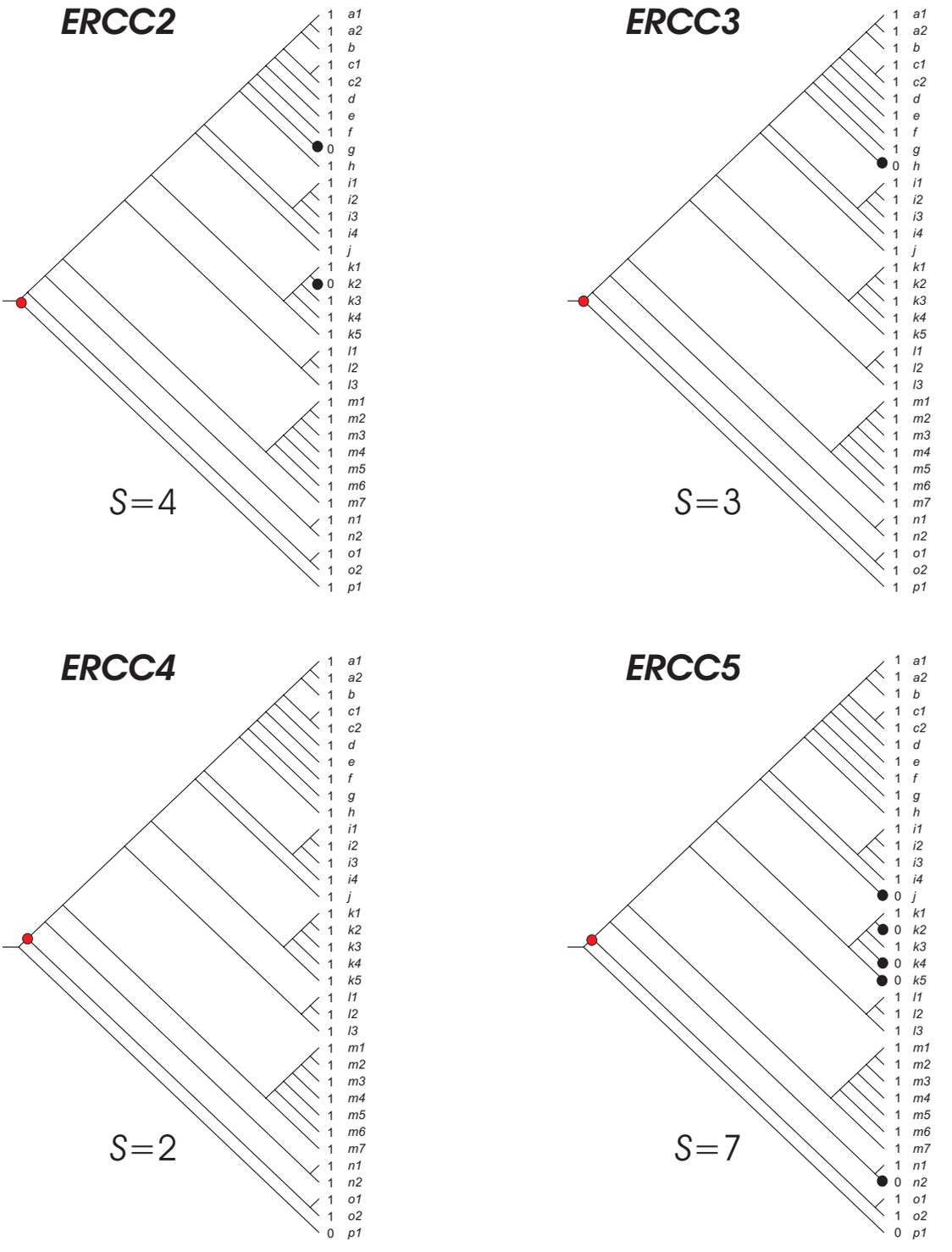
Supplementary Figure S59. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



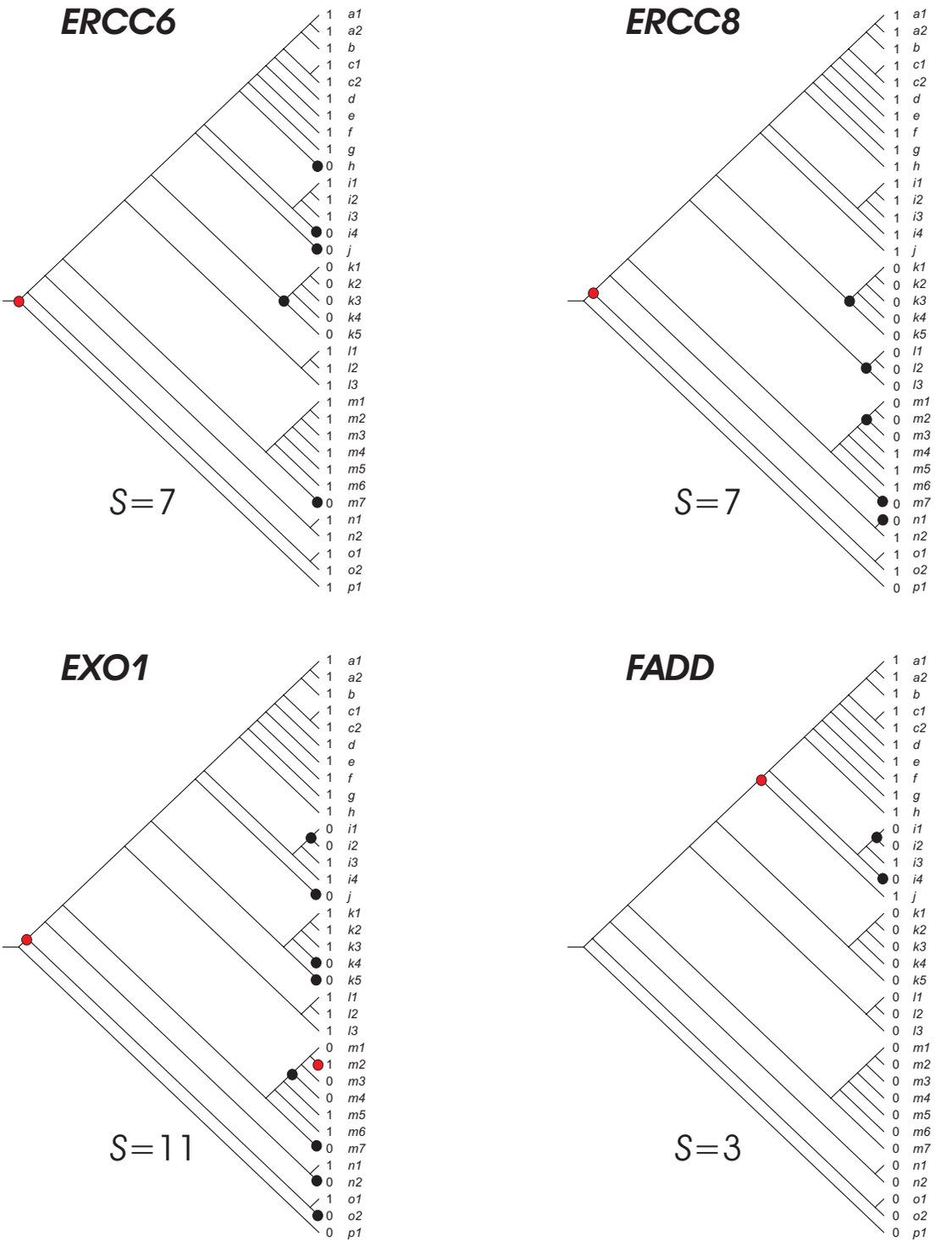
Supplementary Figure S60. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



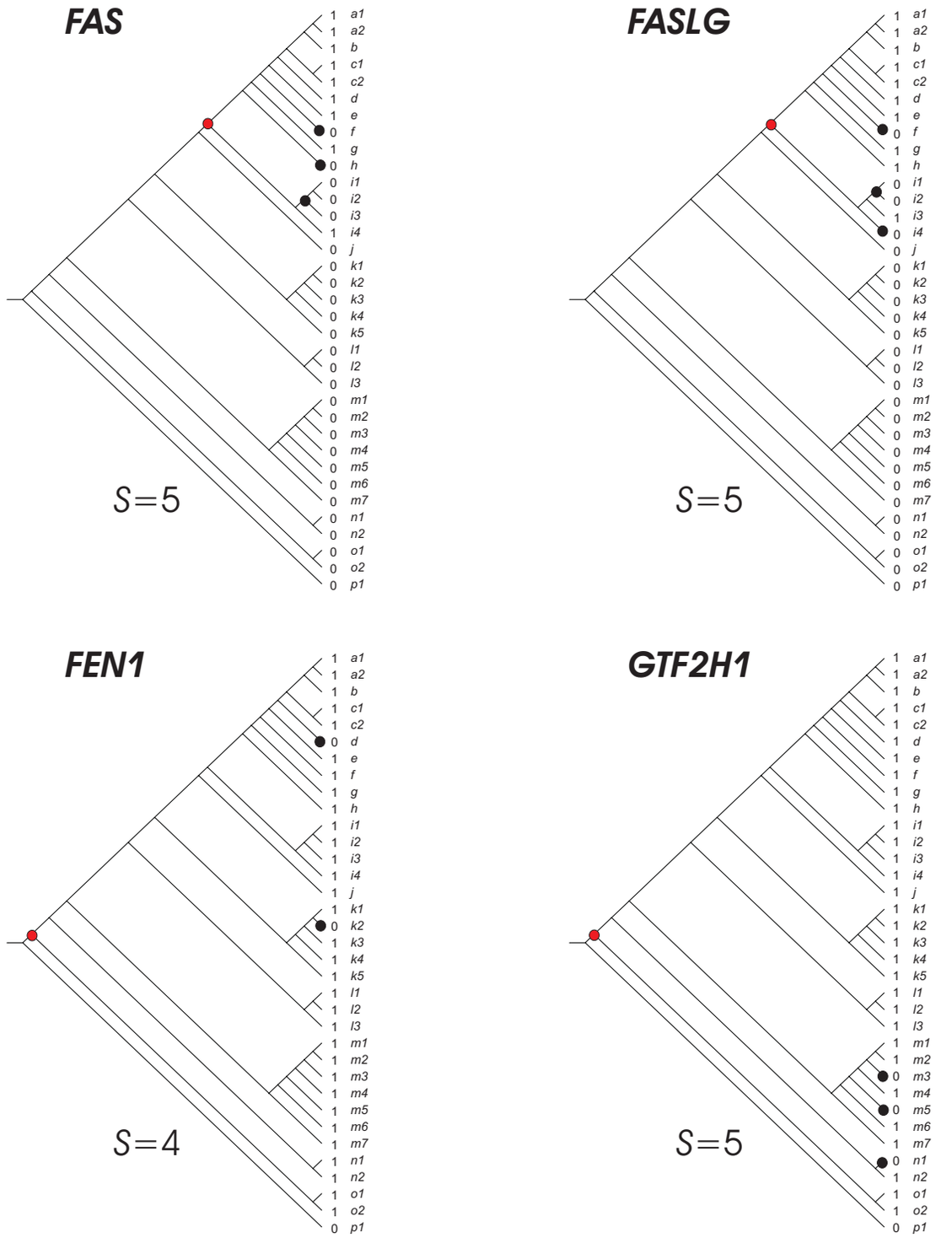
Supplementary Figure S61. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



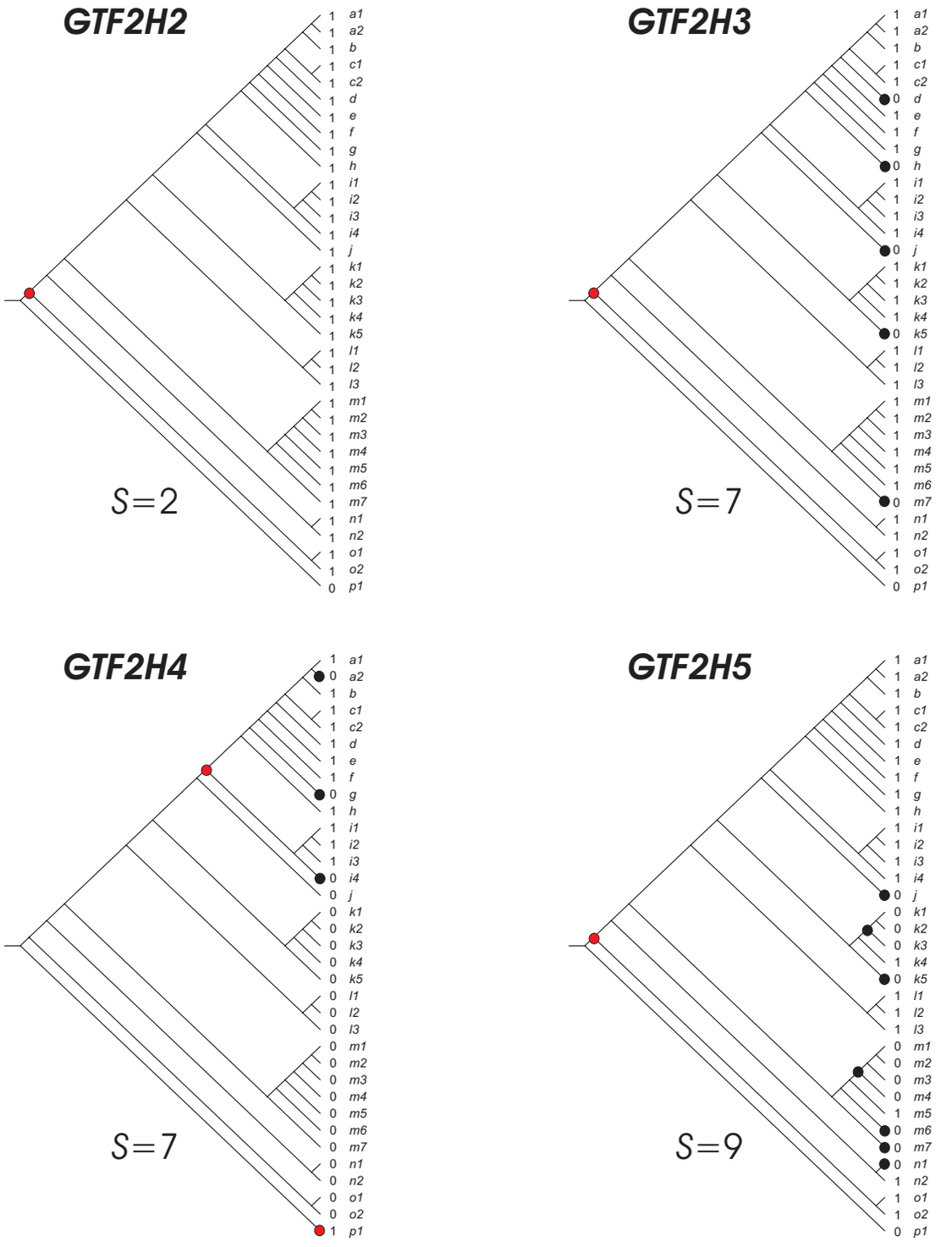
Supplementary Figure S62. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



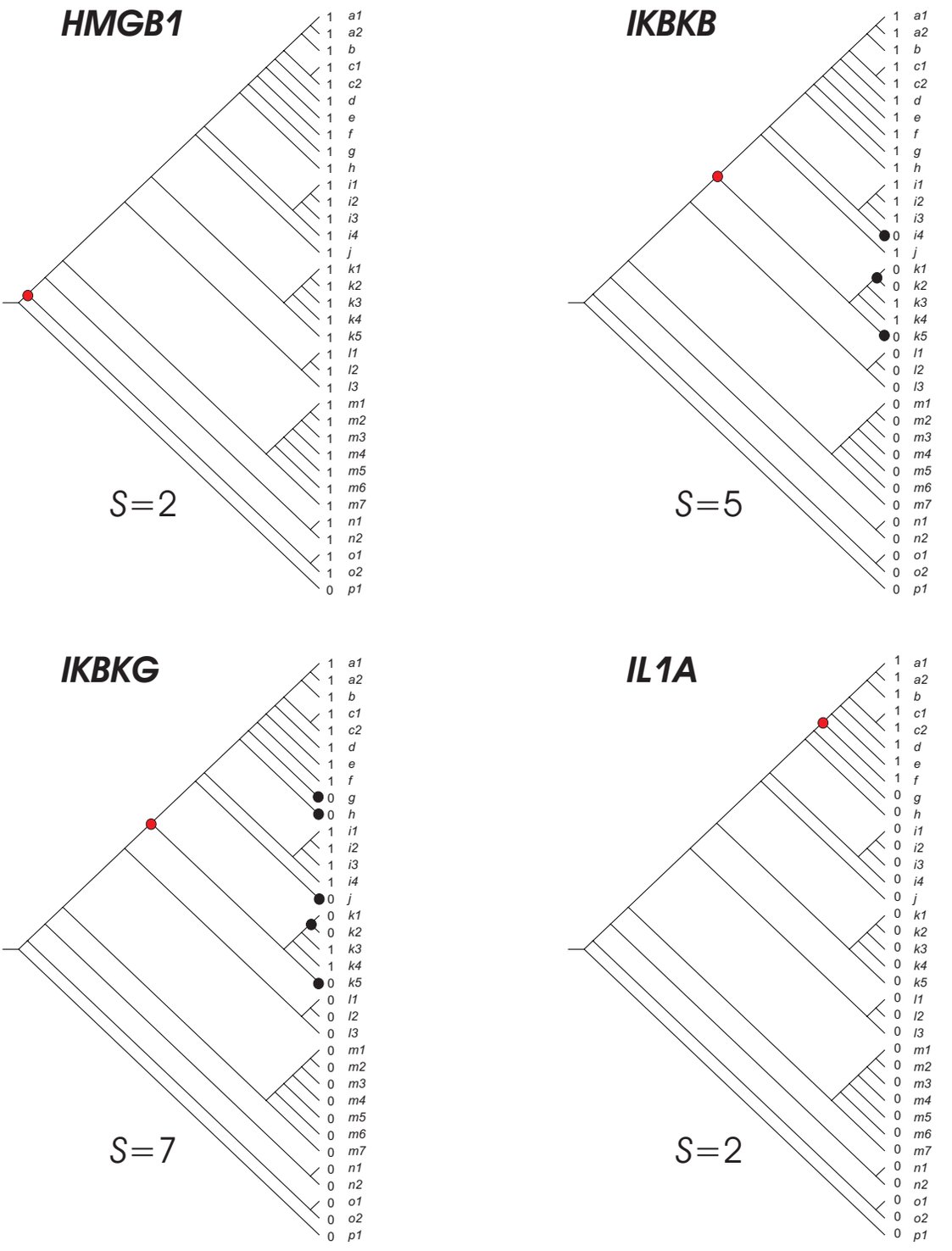
Supplementary Figure S63. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



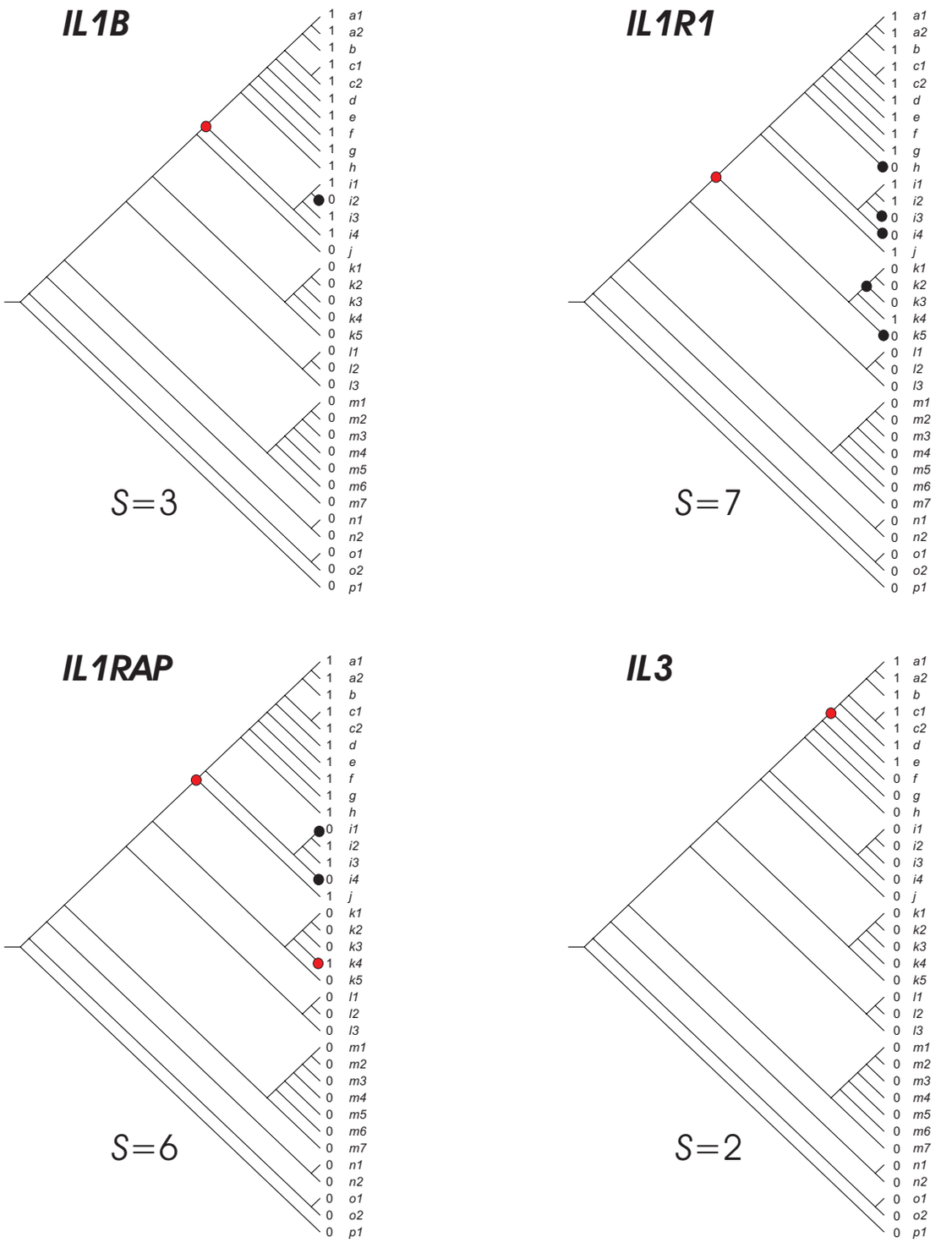
Supplementary Figure S64. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



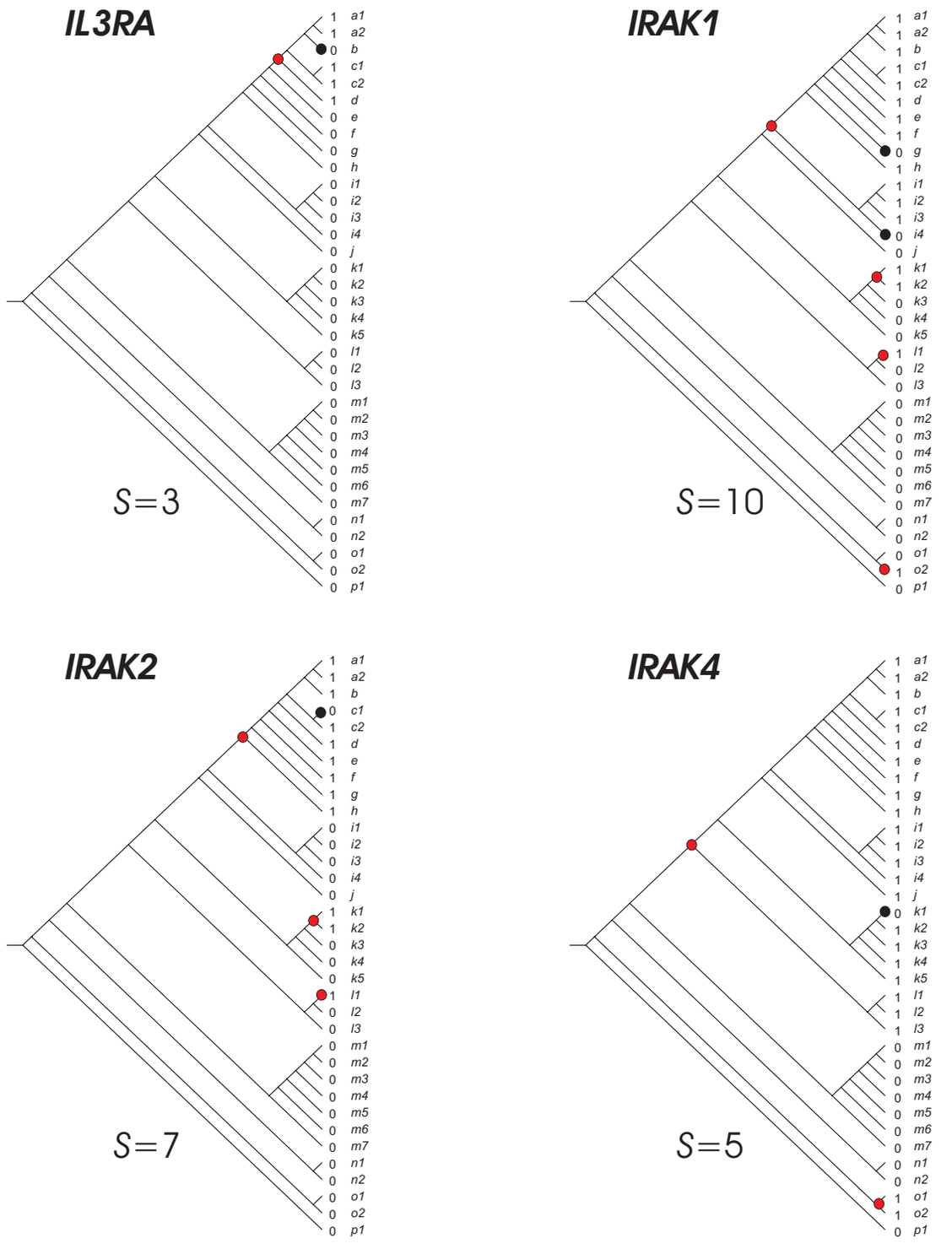
Supplementary Figure S65. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



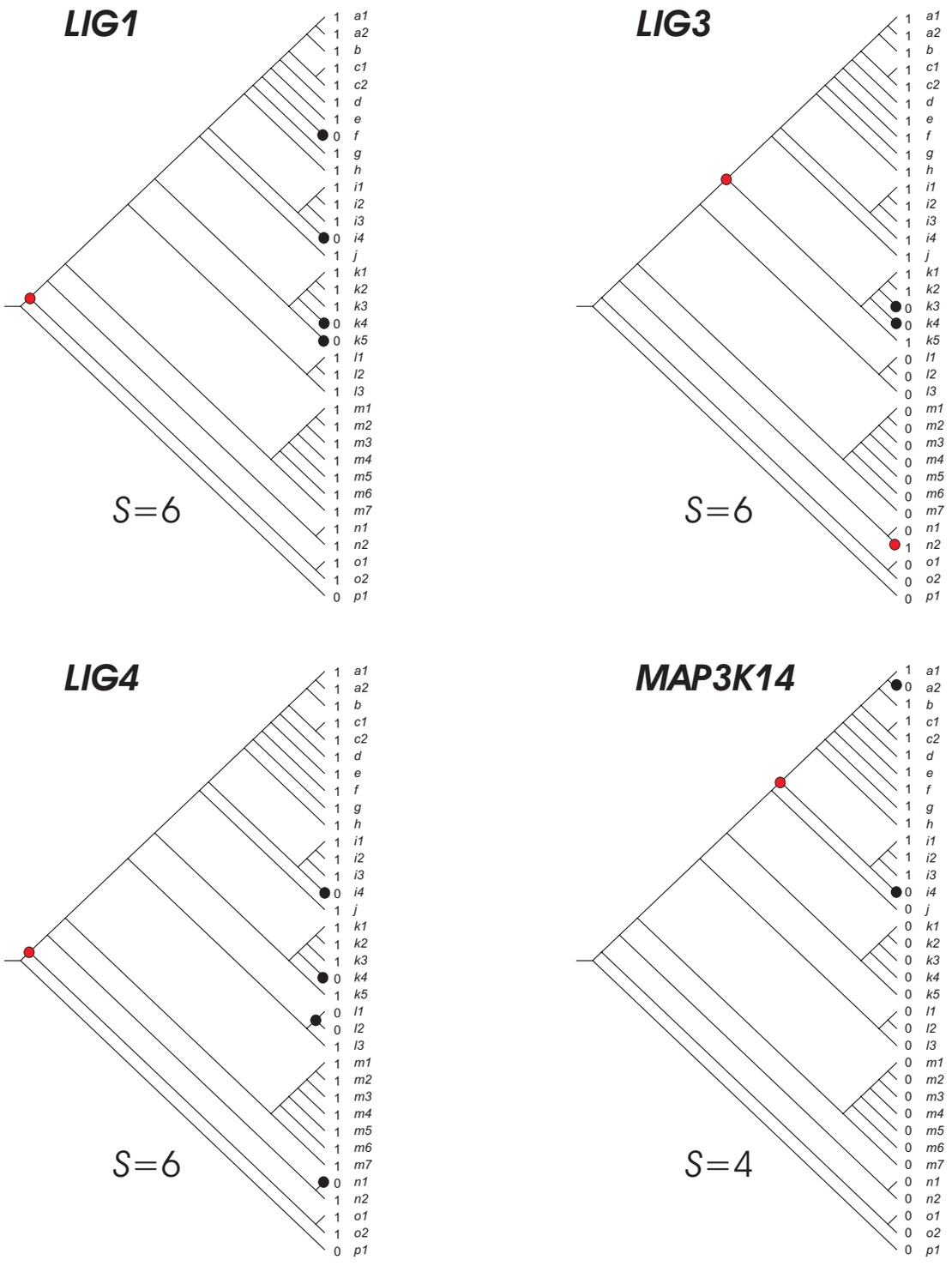
Supplementary Figure S66. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



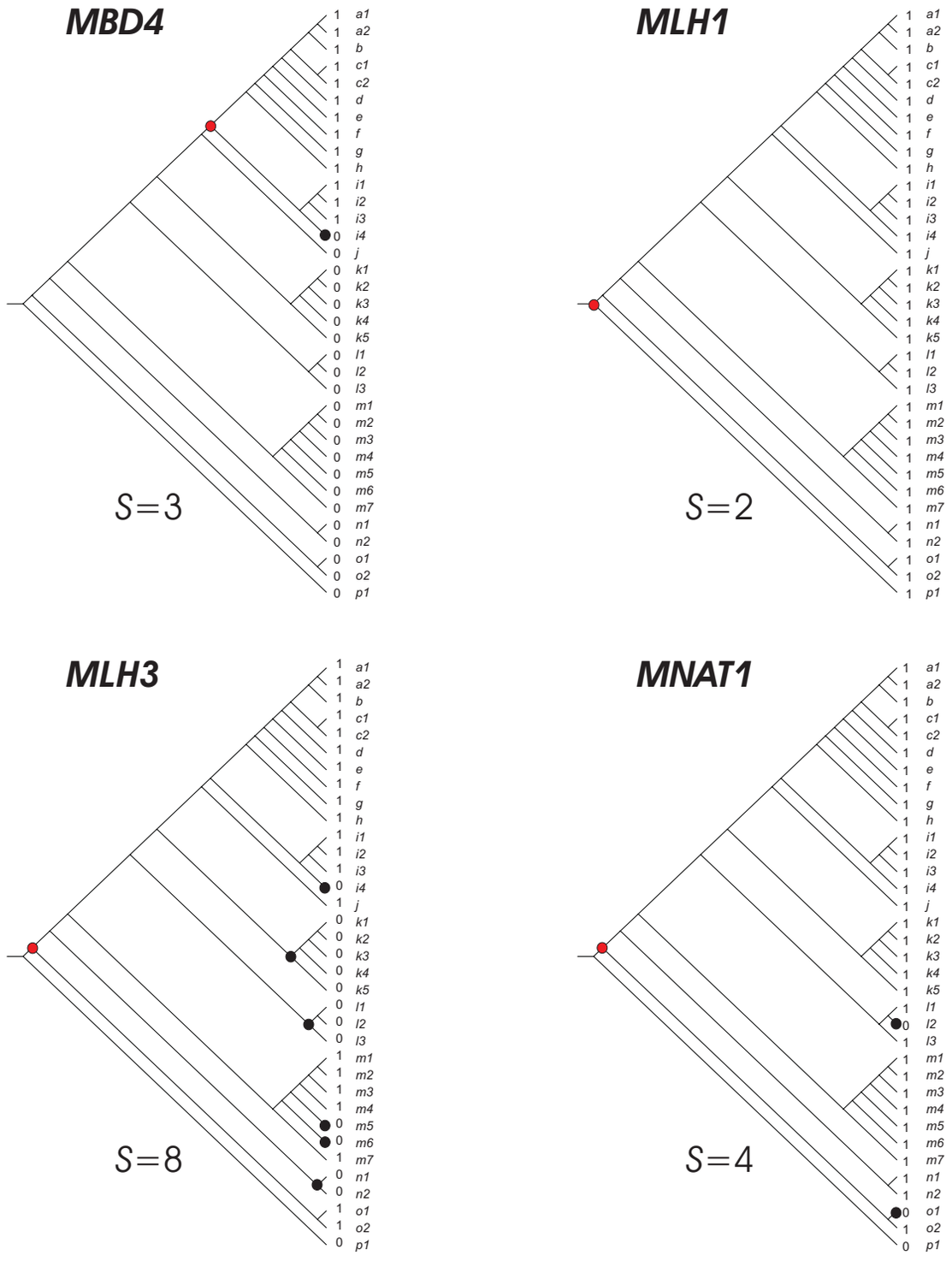
Supplementary Figure S67. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



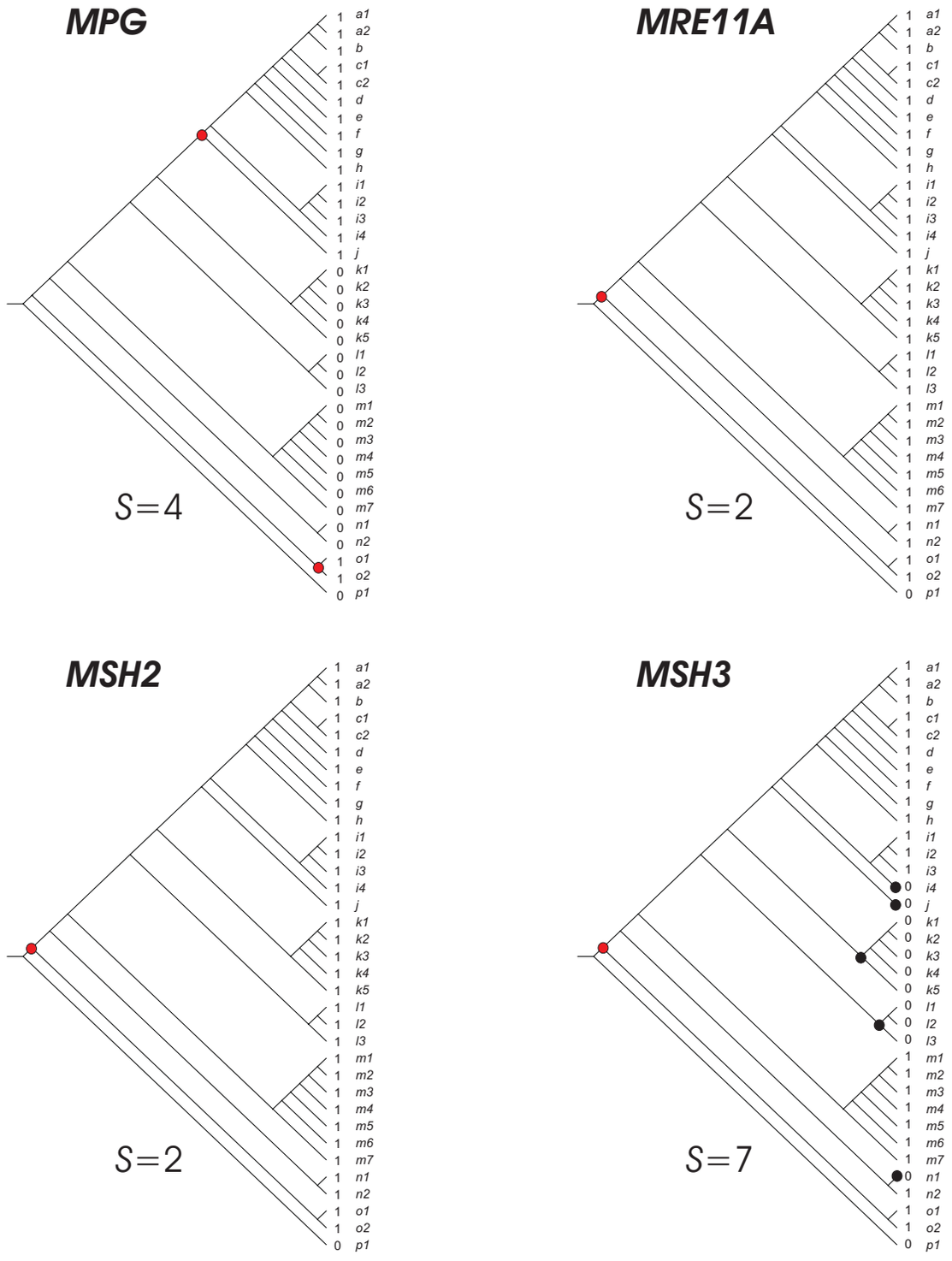
Supplementary Figure S68. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



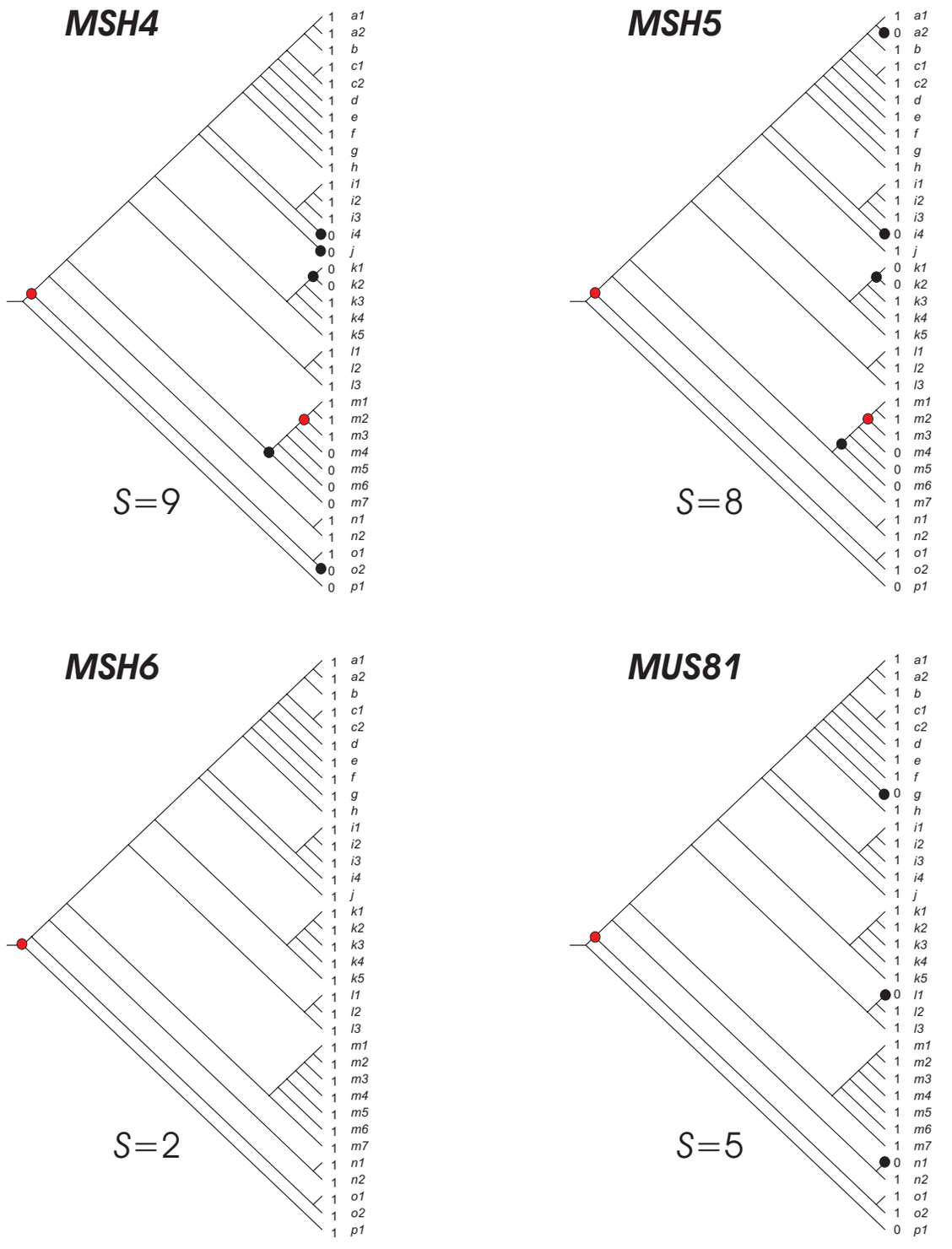
Supplementary Figure S69. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



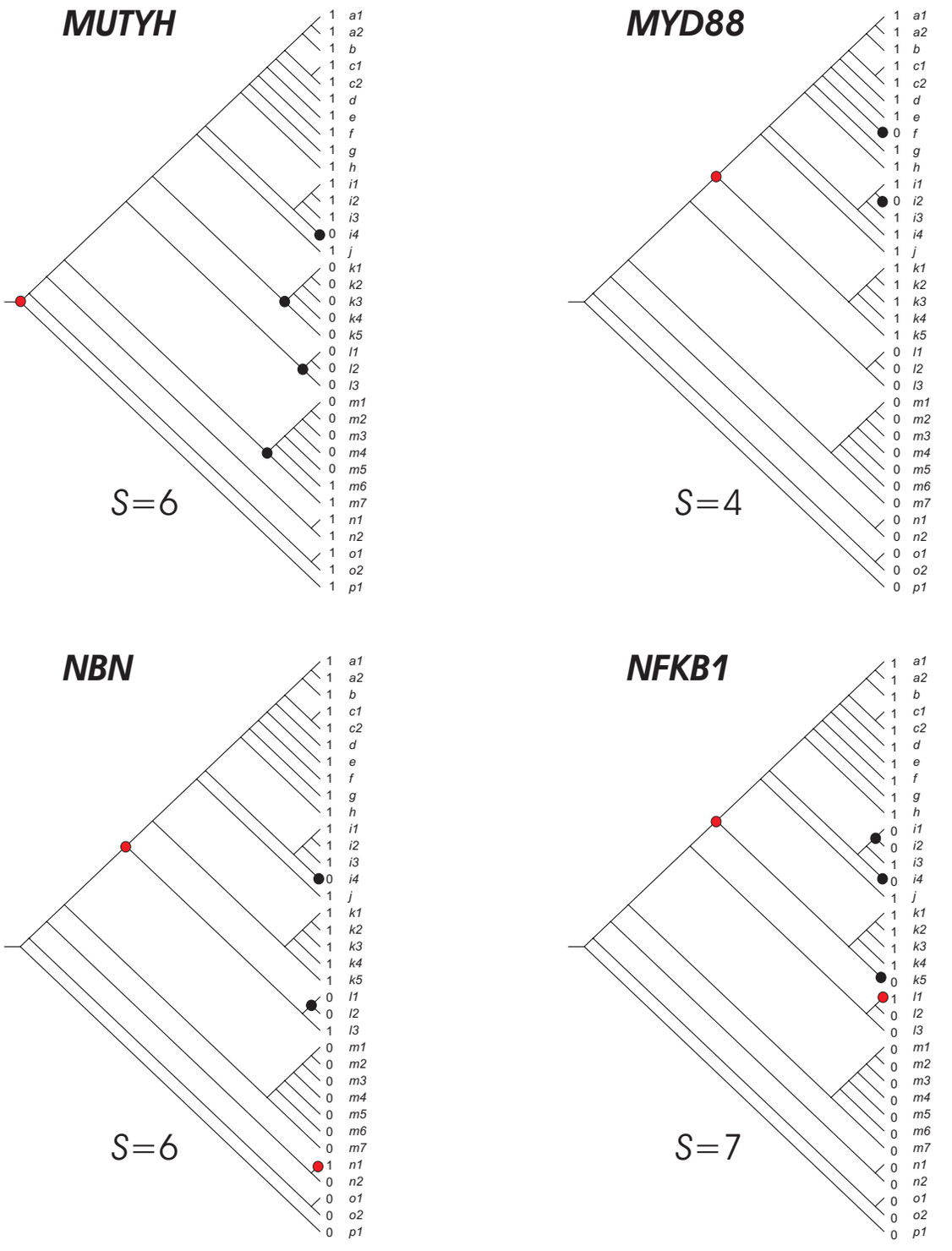
Supplementary Figure S70. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



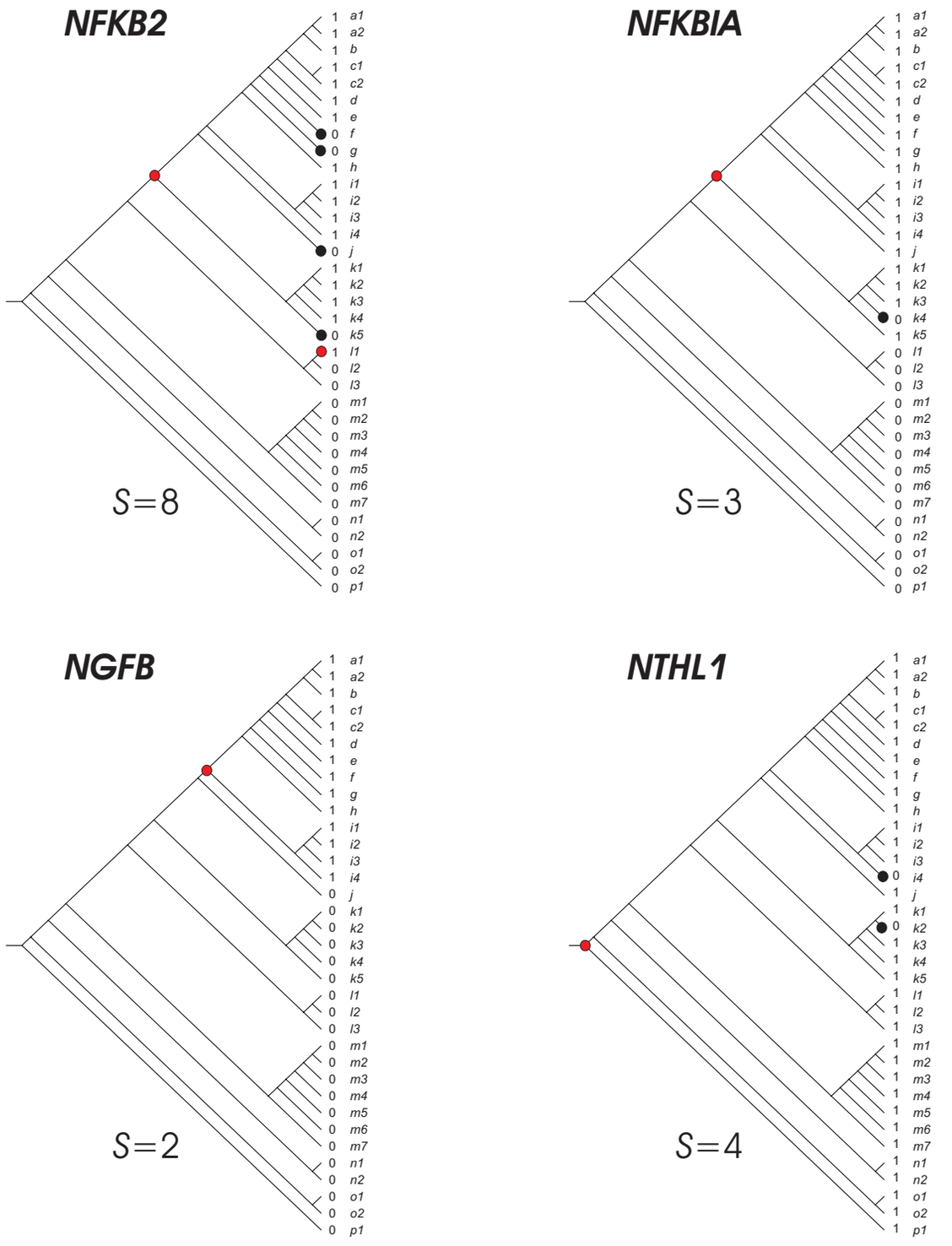
Supplementary Figure S71. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



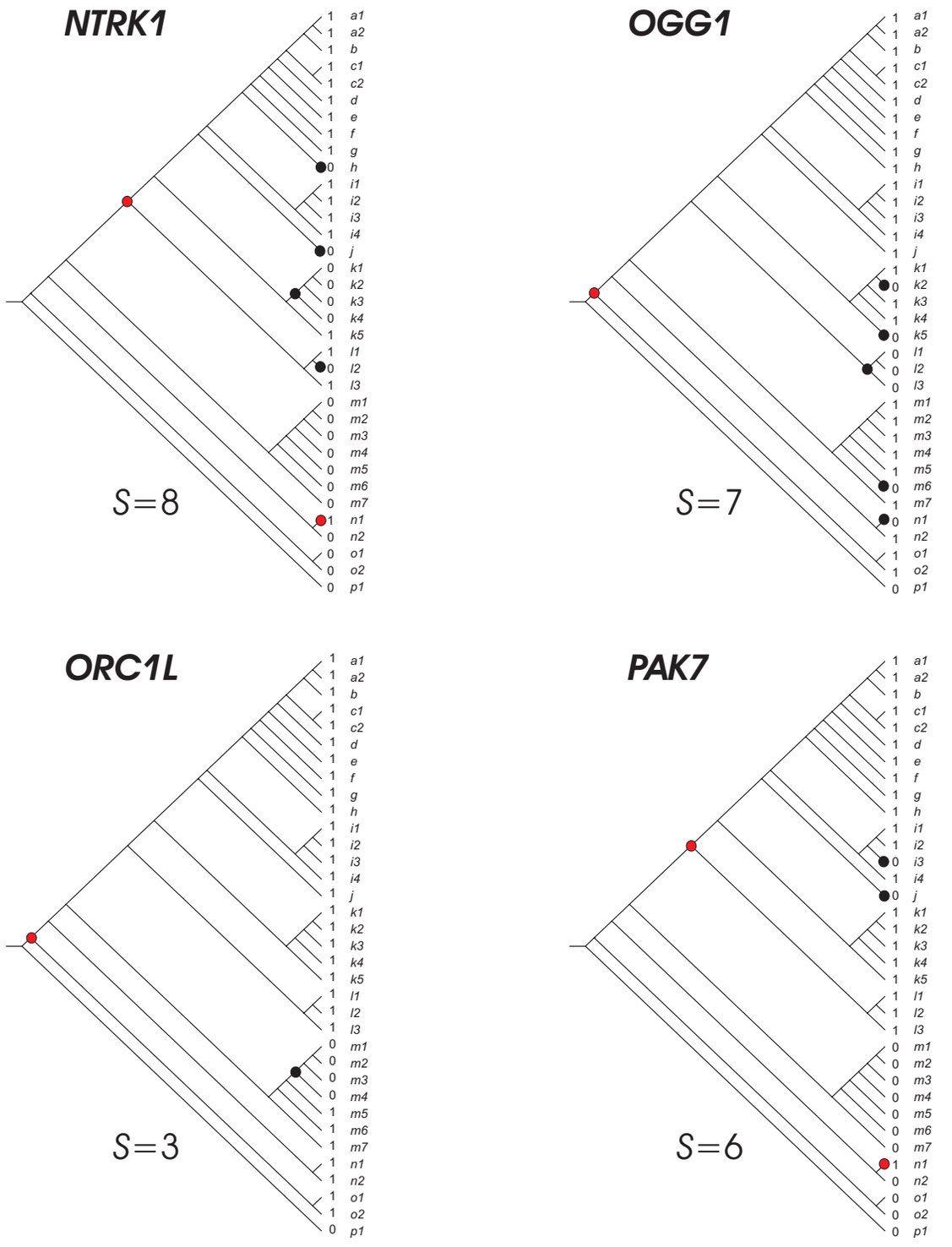
Supplementary Figure S72. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



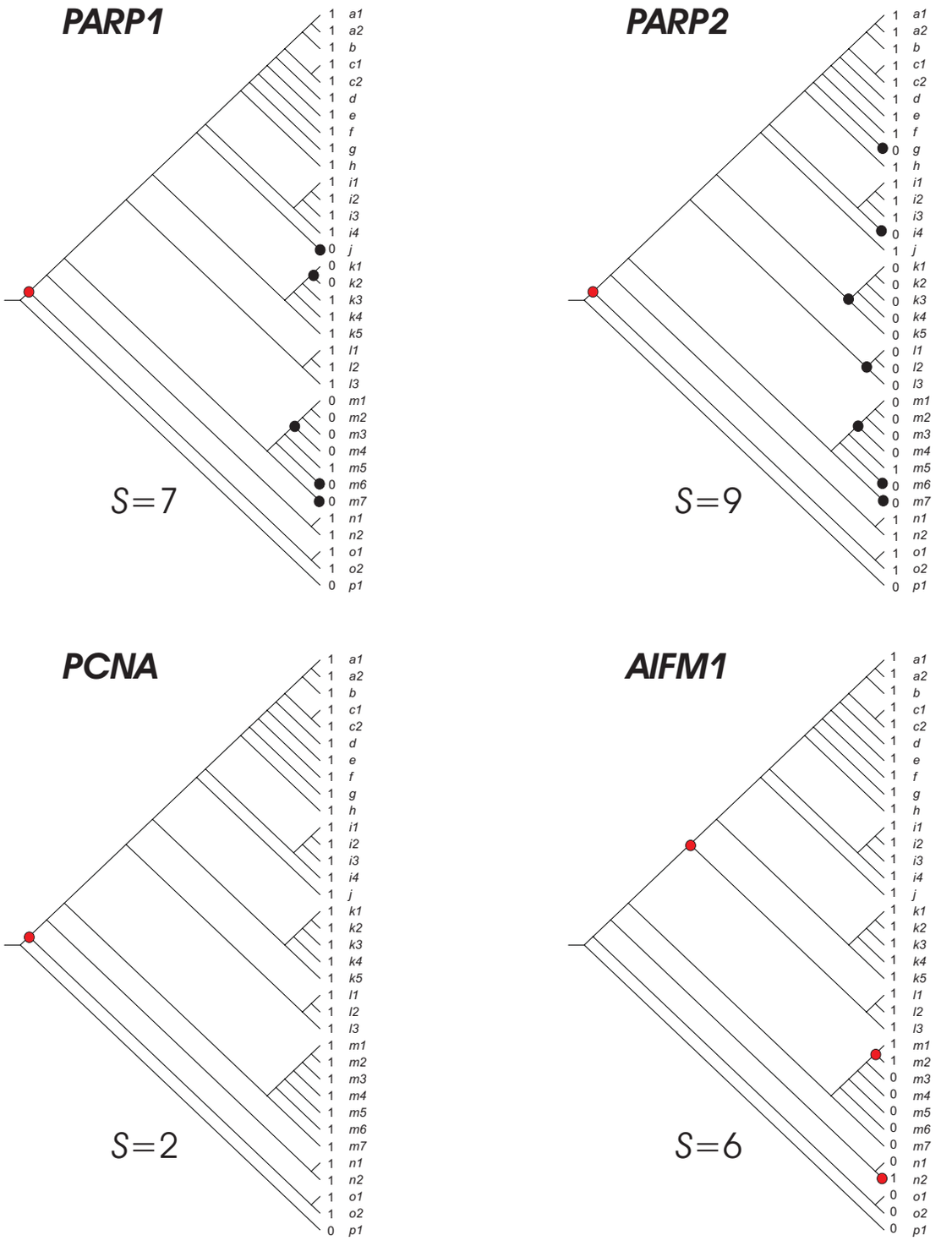
Supplementary Figure S73. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



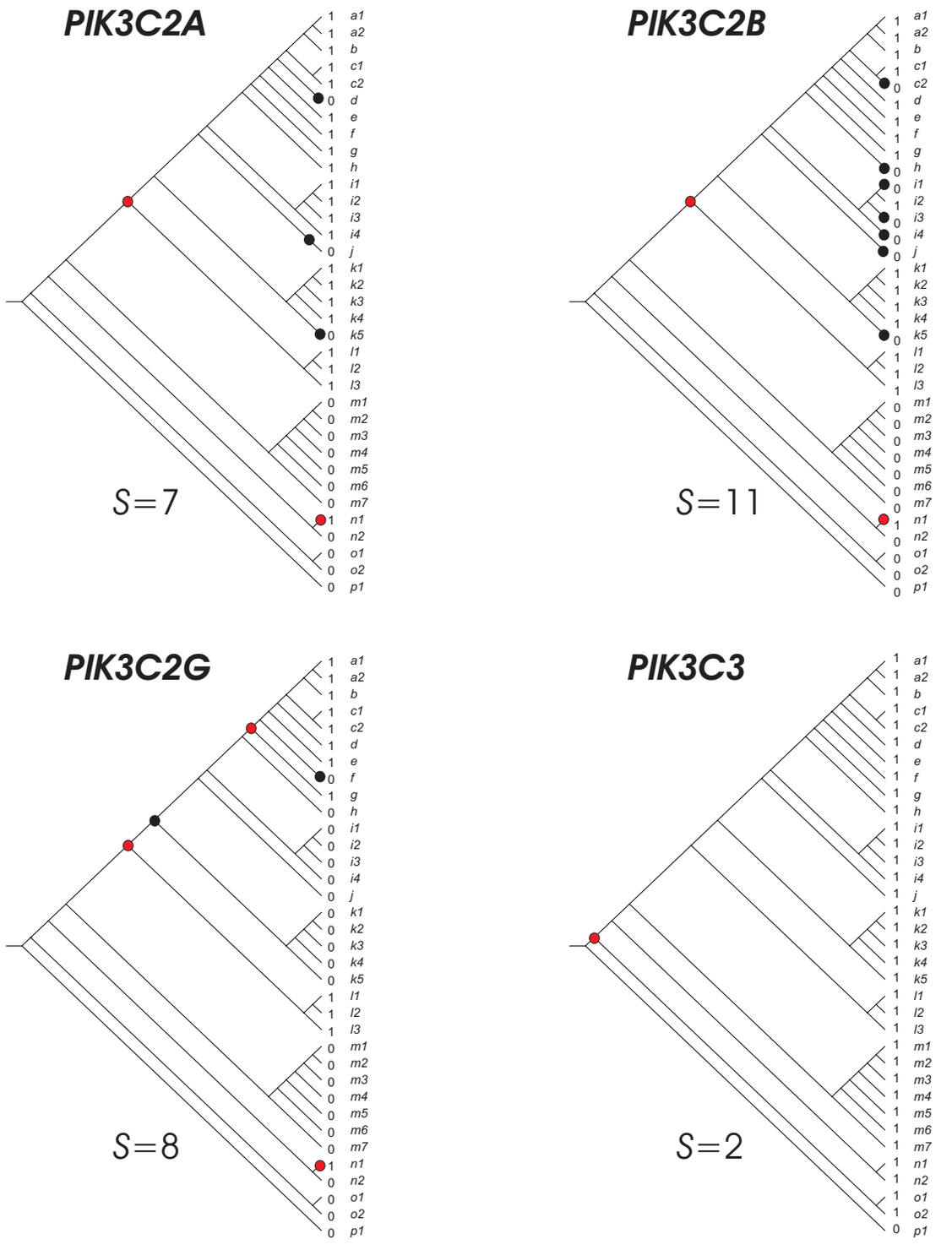
Supplementary Figure S74. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



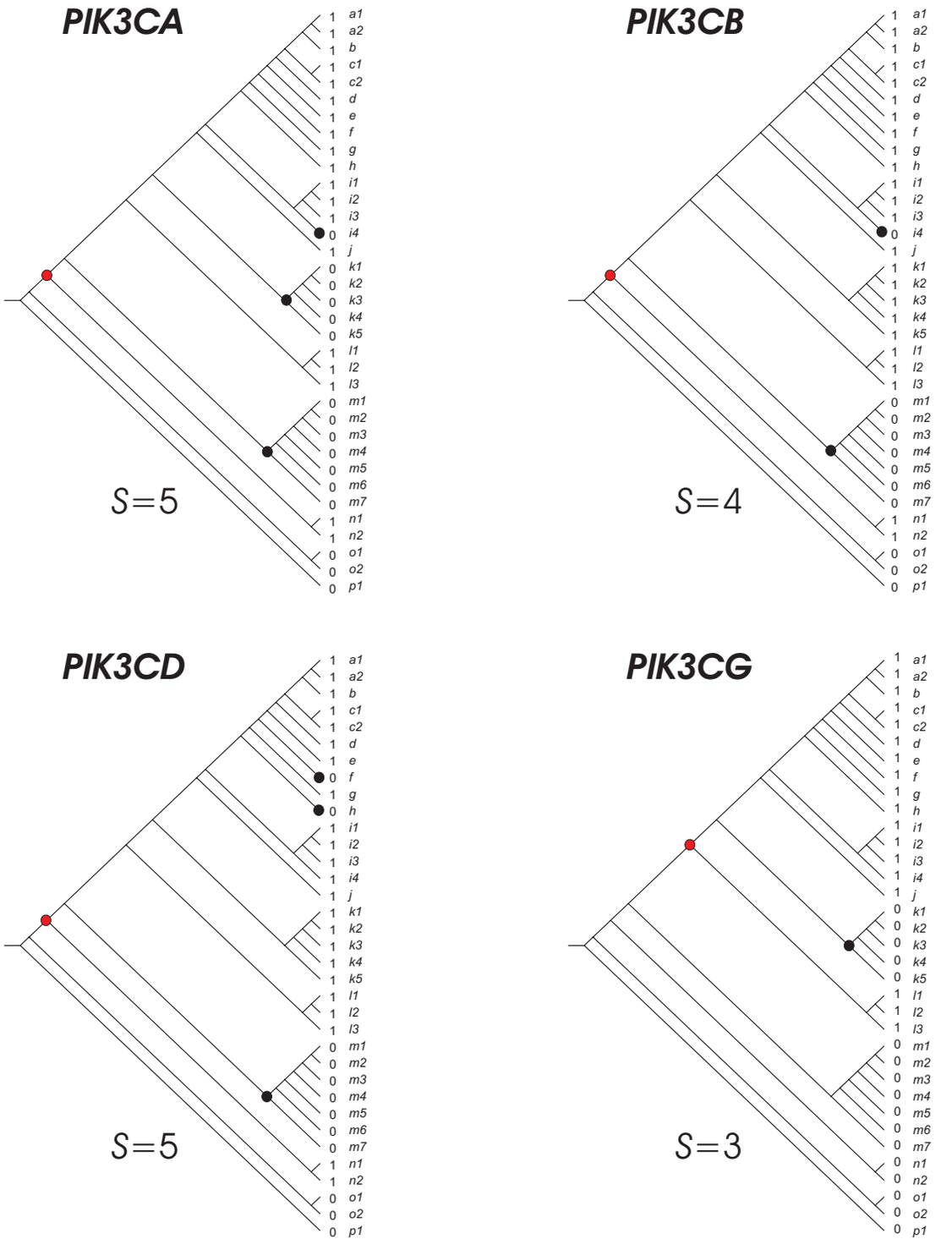
Supplementary Figure S75. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



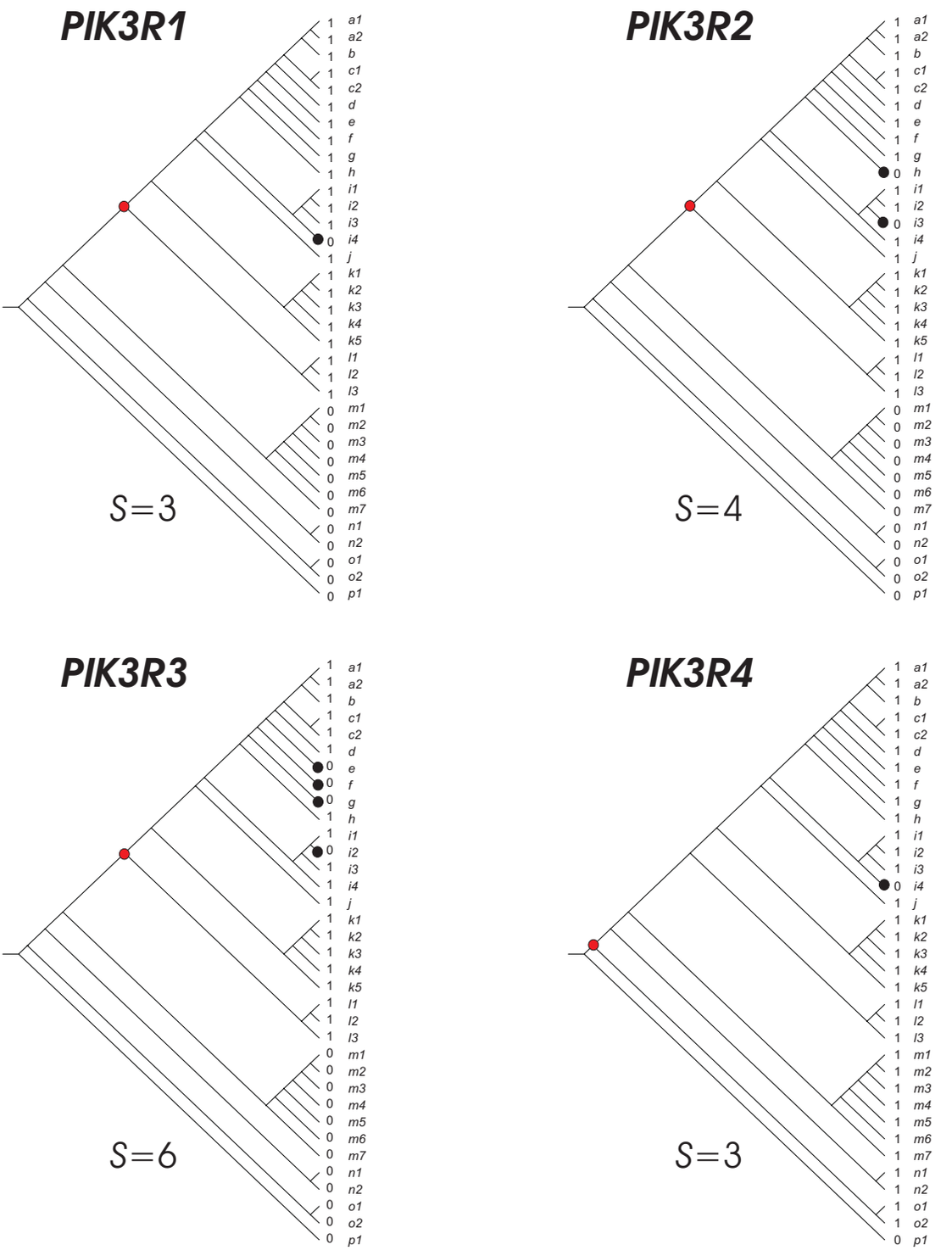
Supplementary Figure S76. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



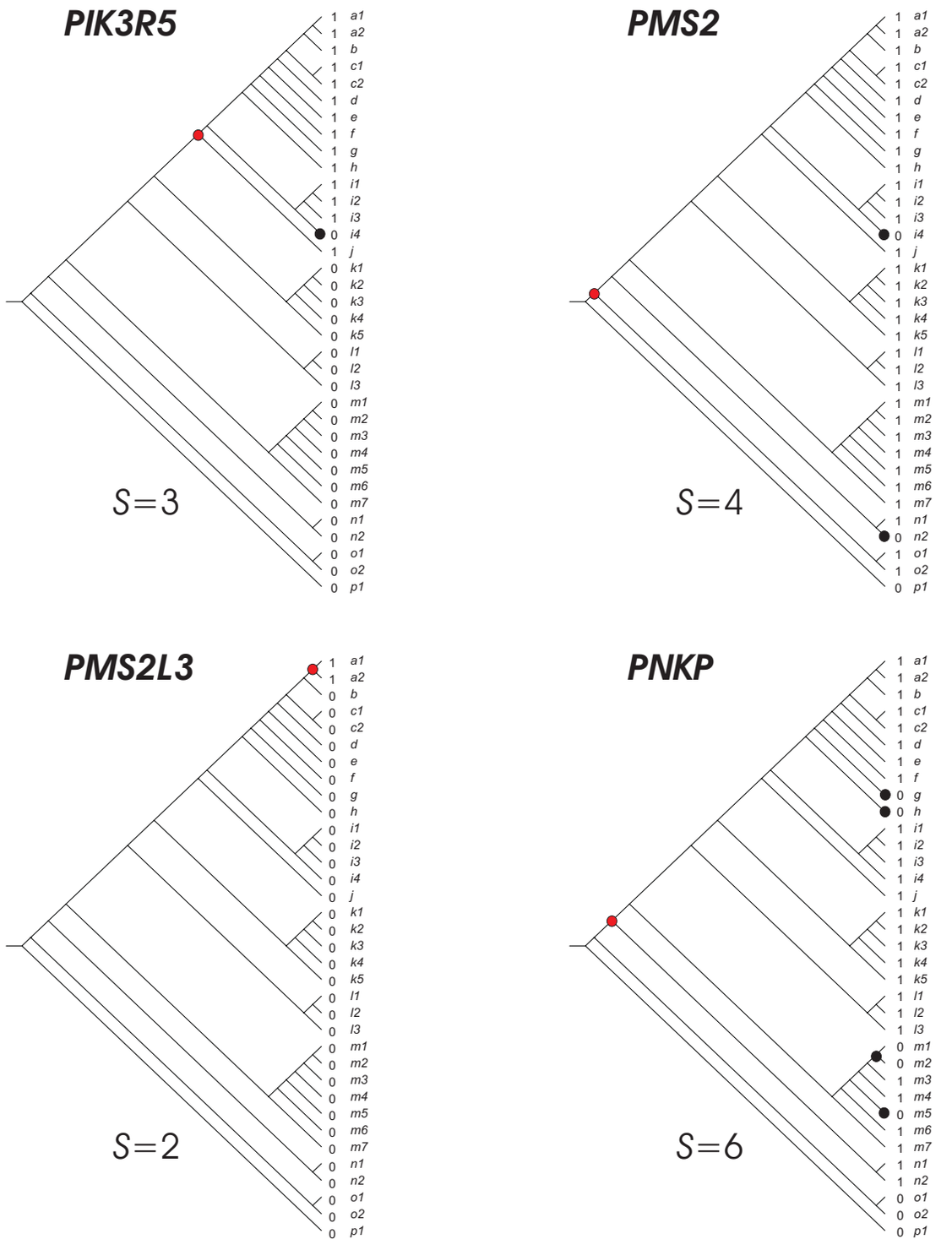
Supplementary Figure S77. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



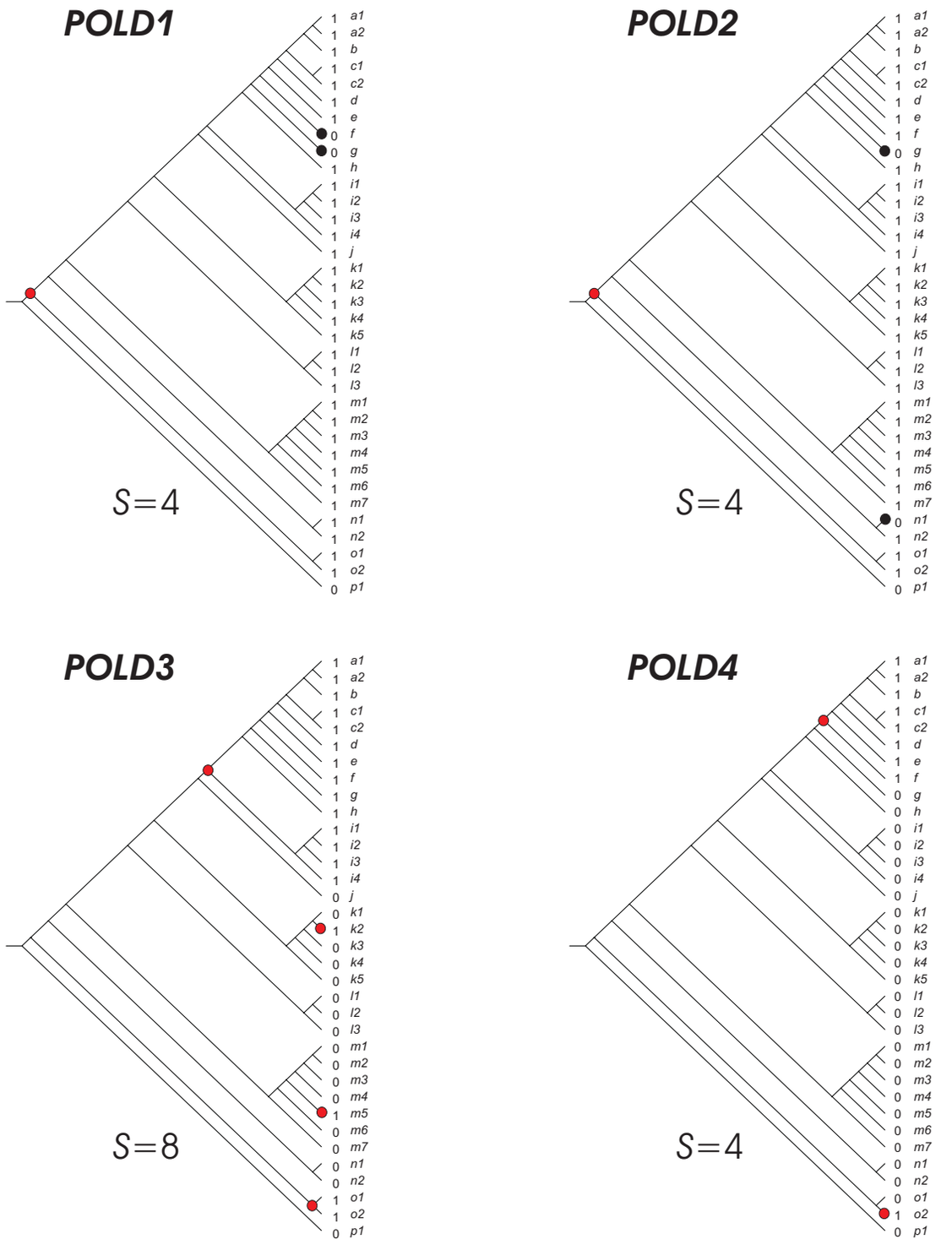
Supplementary Figure S78. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



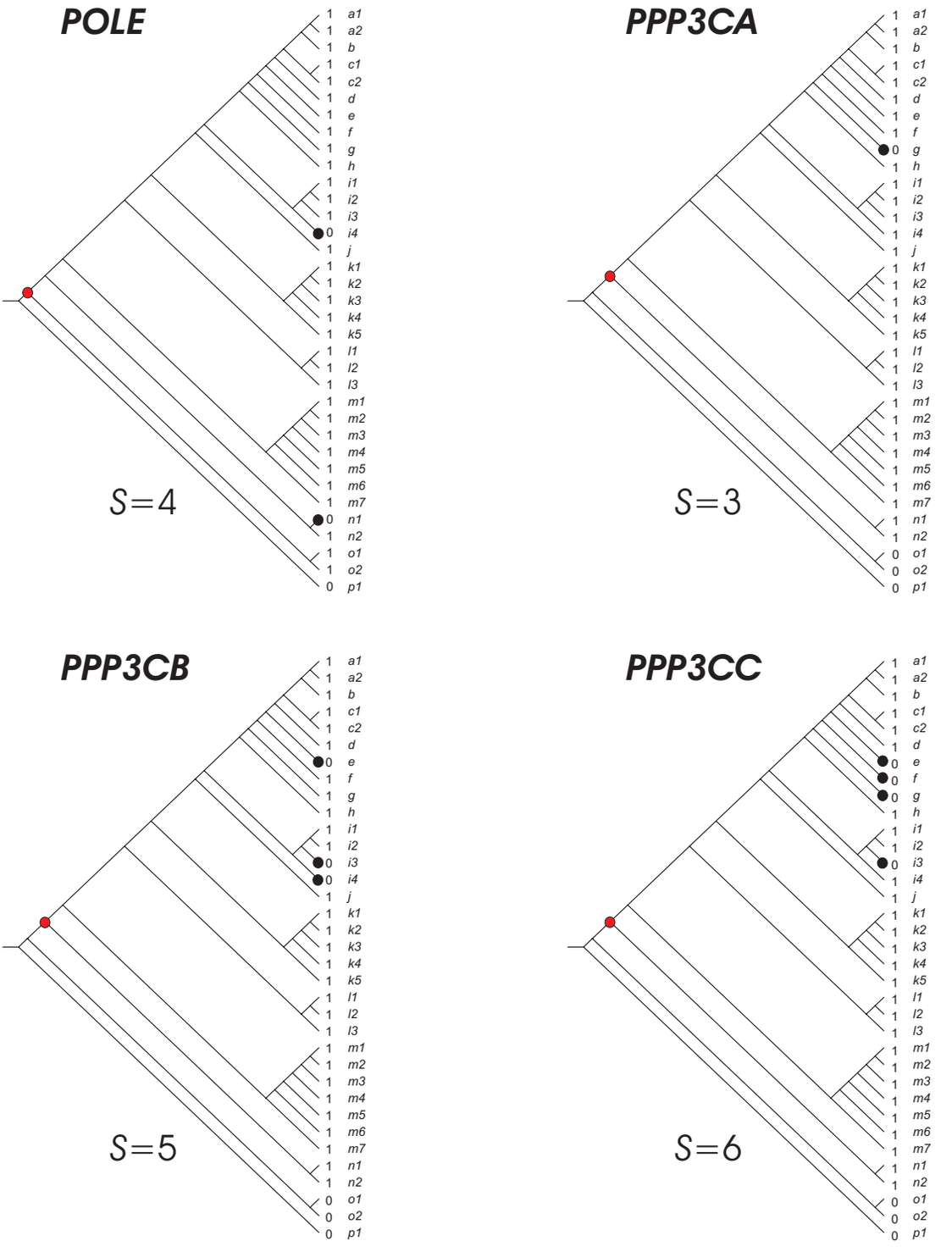
Supplementary Figure S79. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



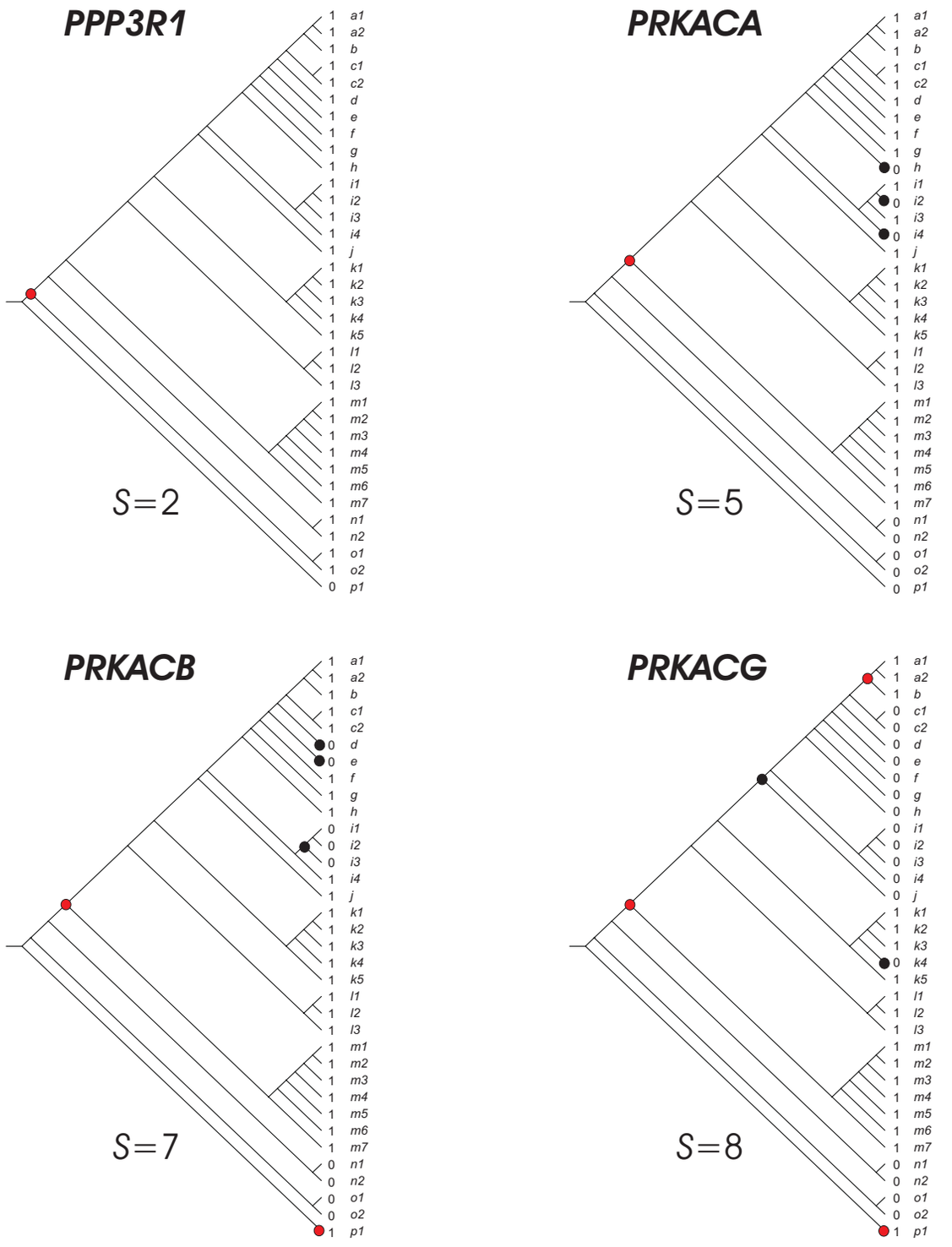
Supplementary Figure S80. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



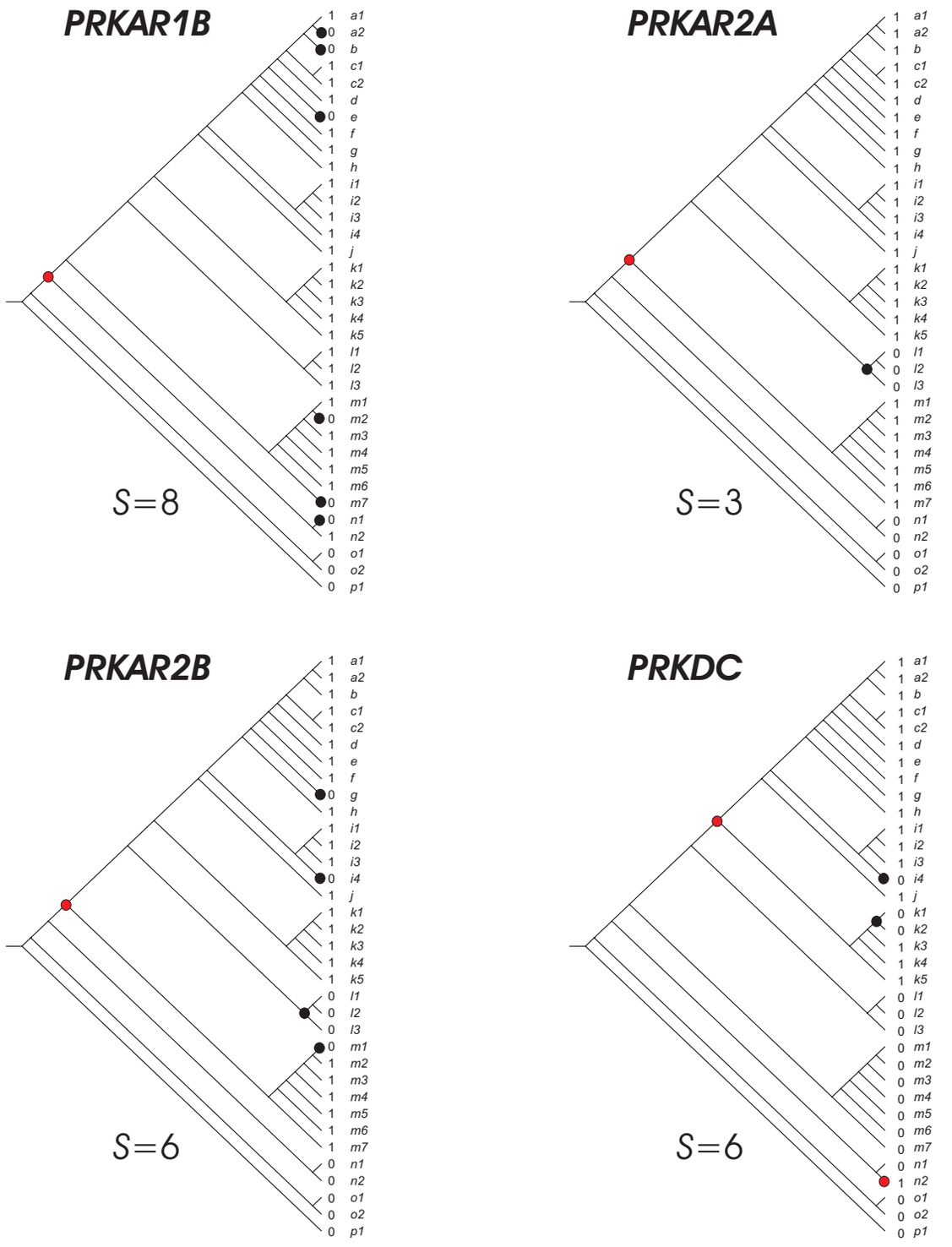
Supplementary Figure S81. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



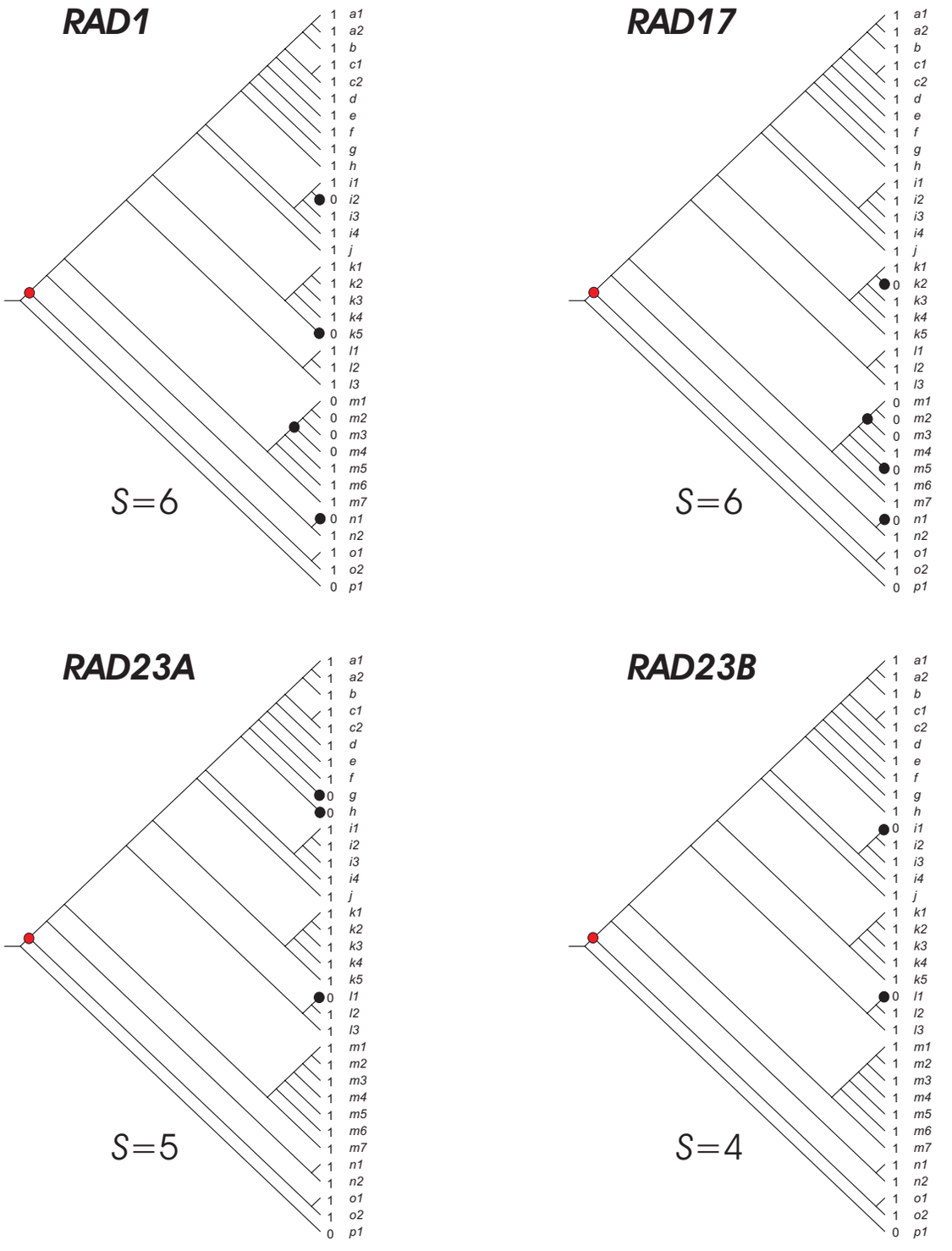
Supplementary Figure S82. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



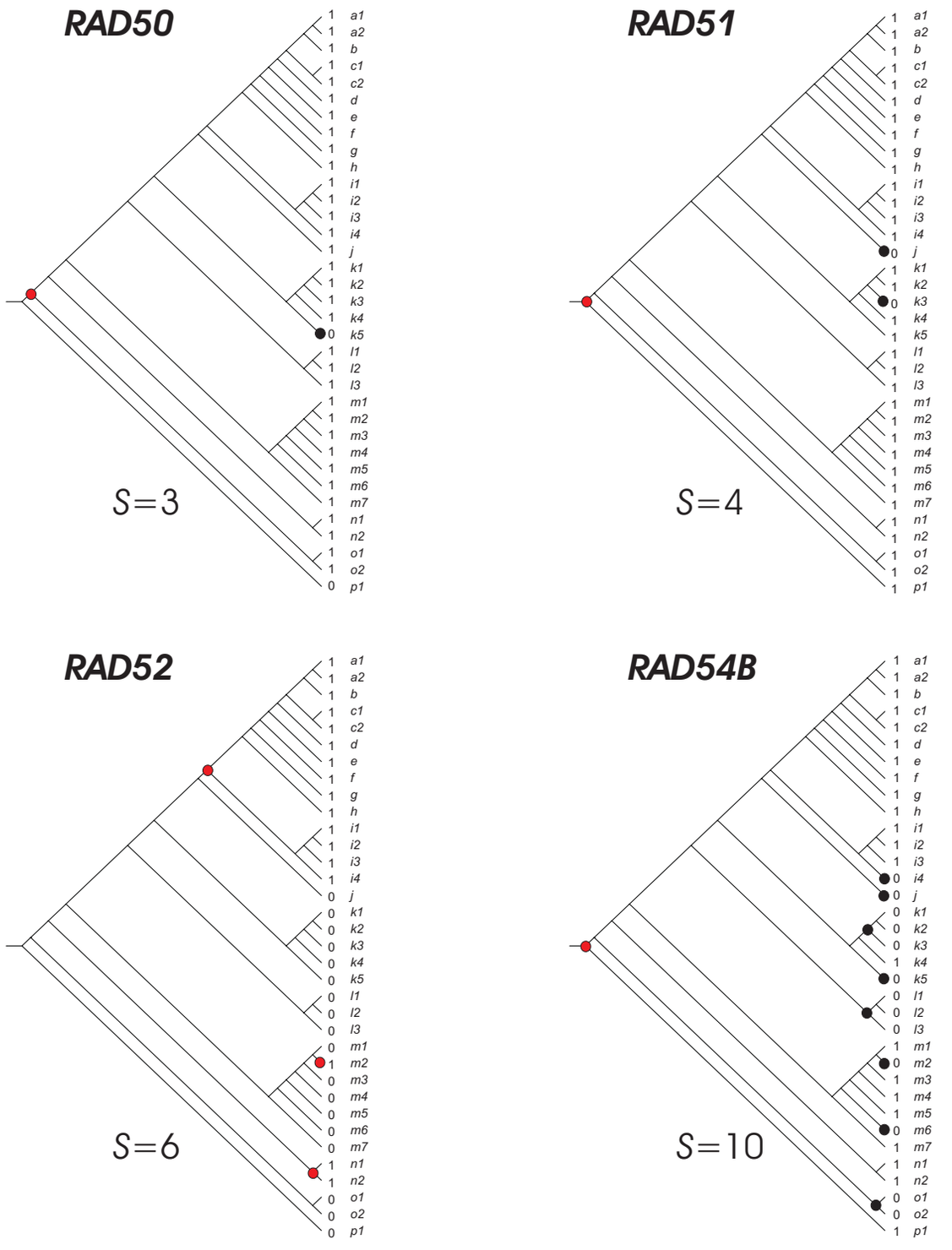
Supplementary Figure S83. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



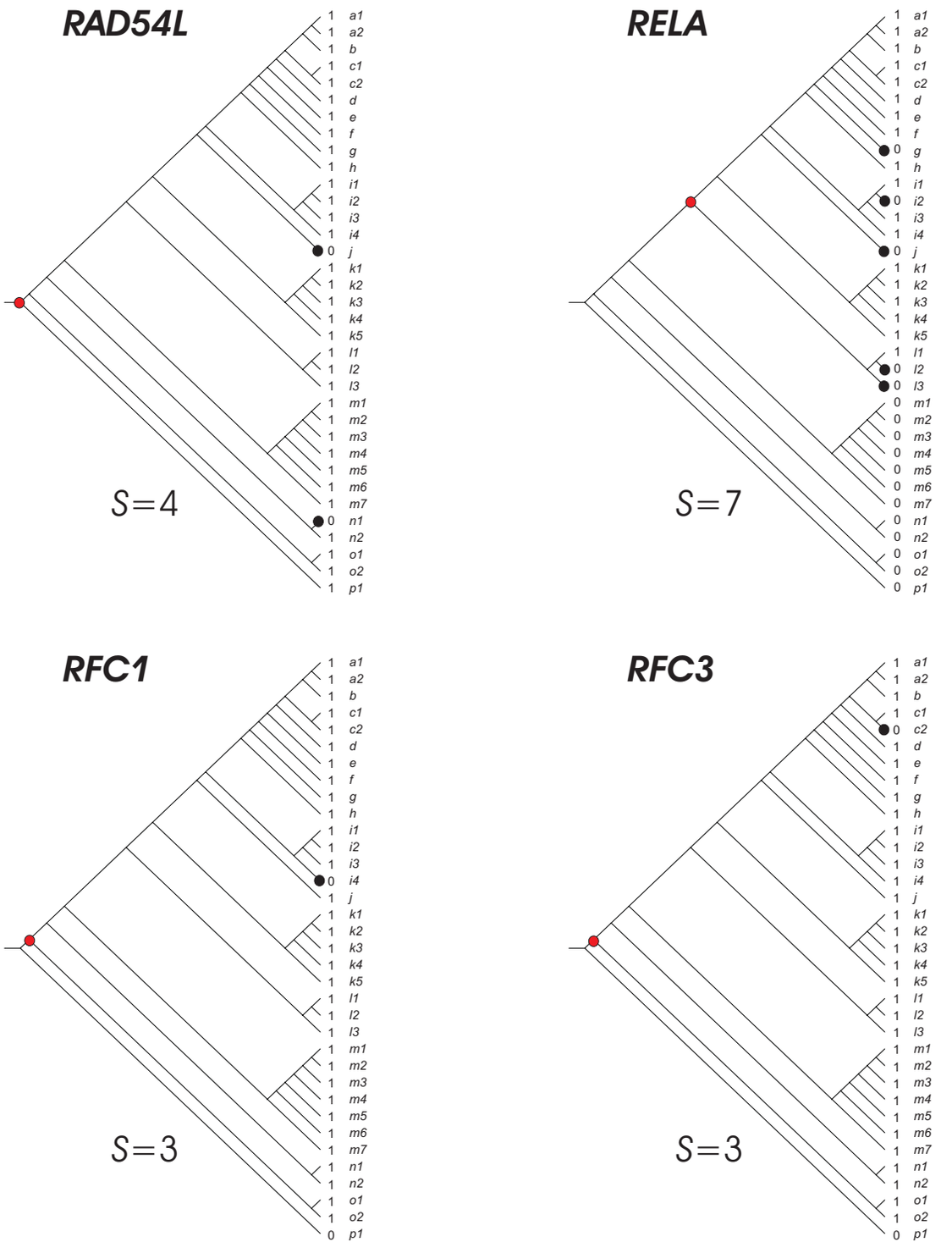
Supplementary Figure S84. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



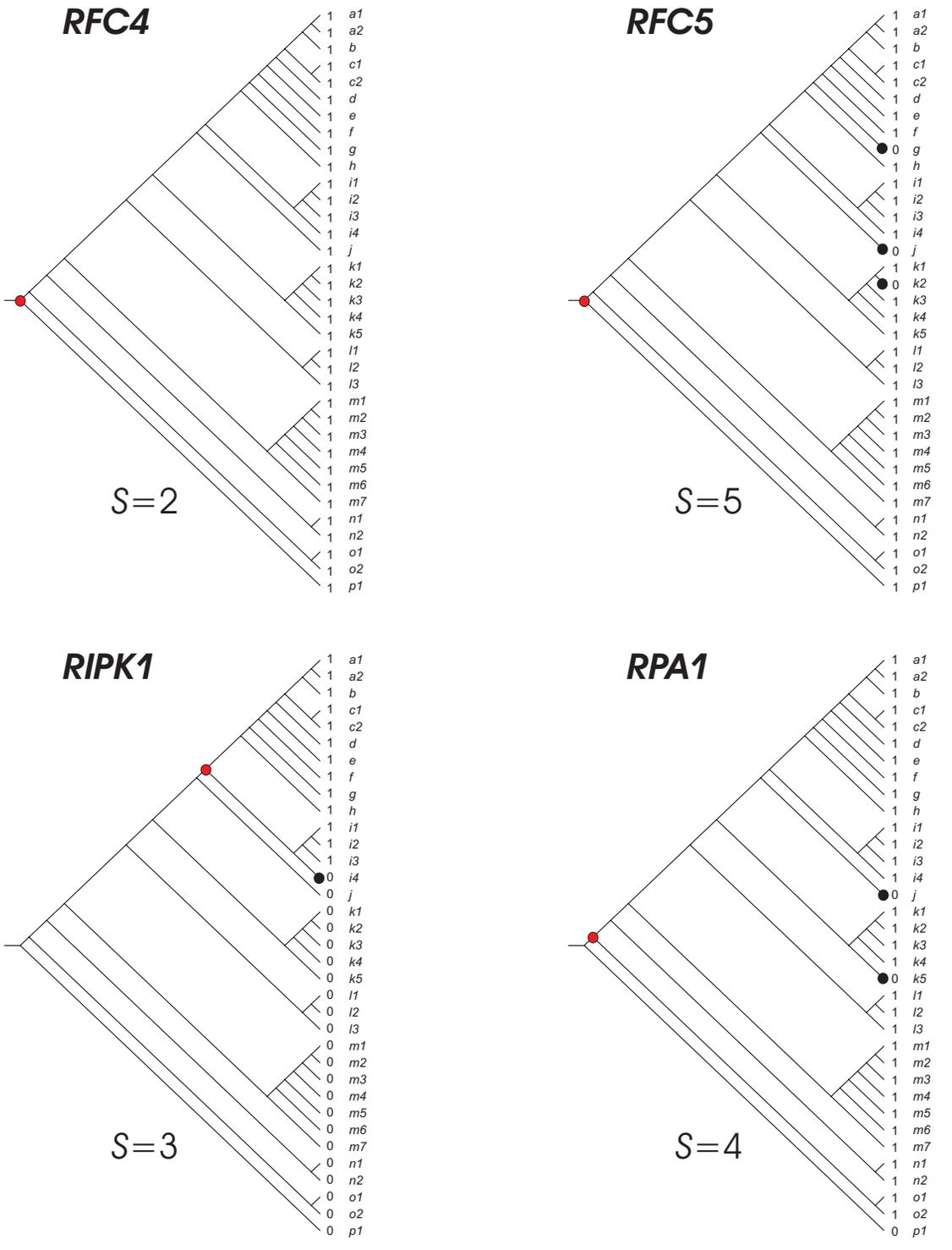
Supplementary Figure S85. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



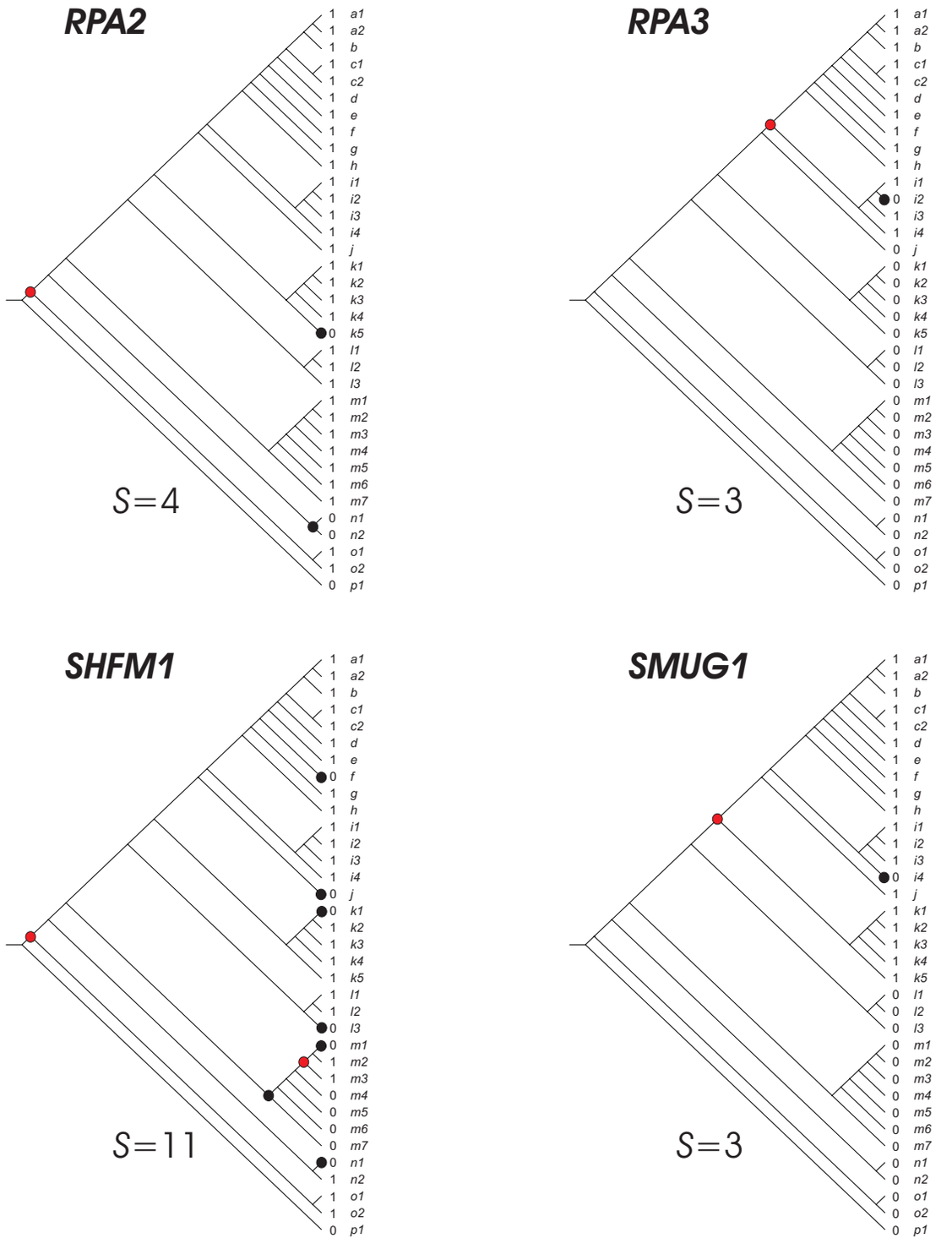
Supplementary Figure S86. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



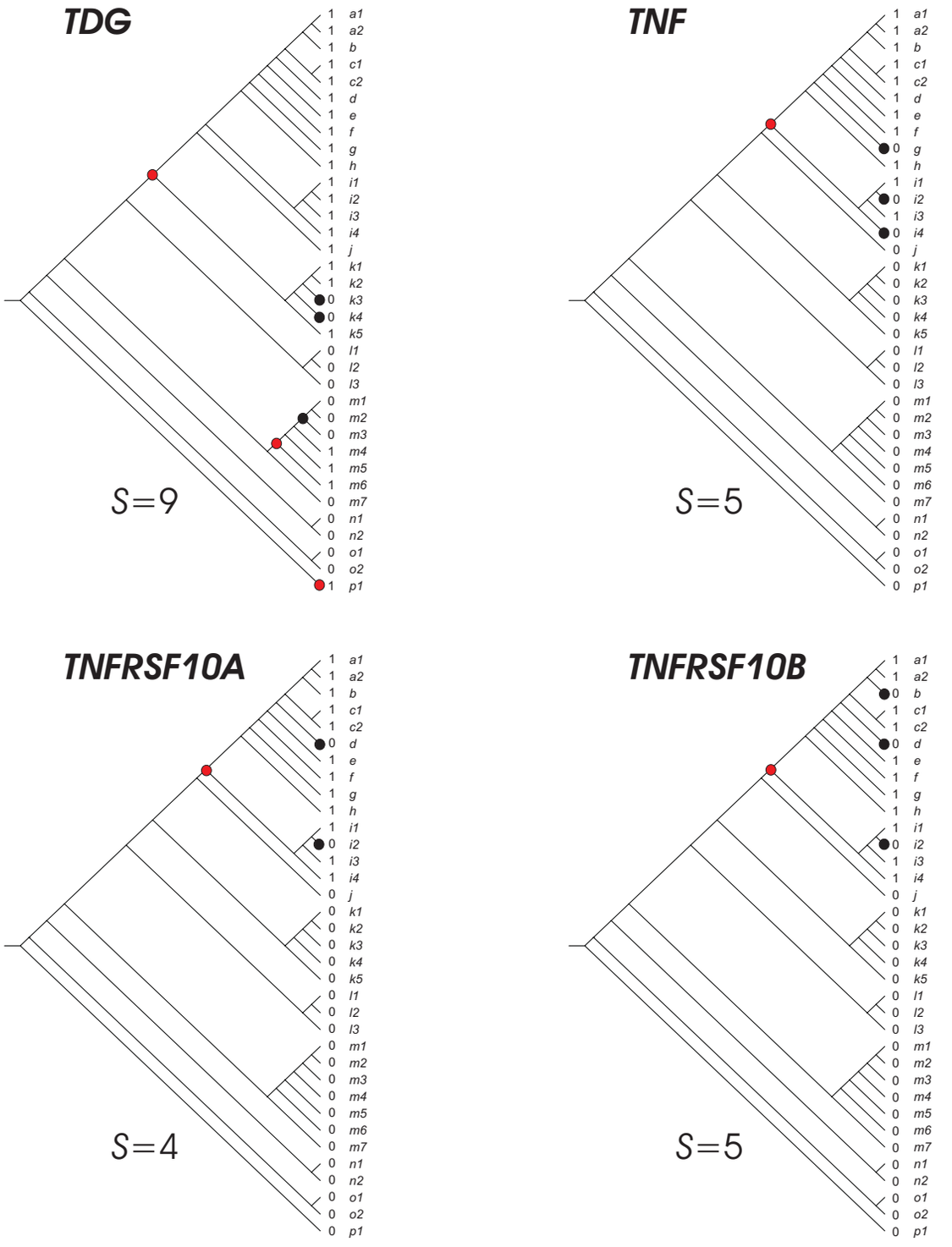
Supplementary Figure S87. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



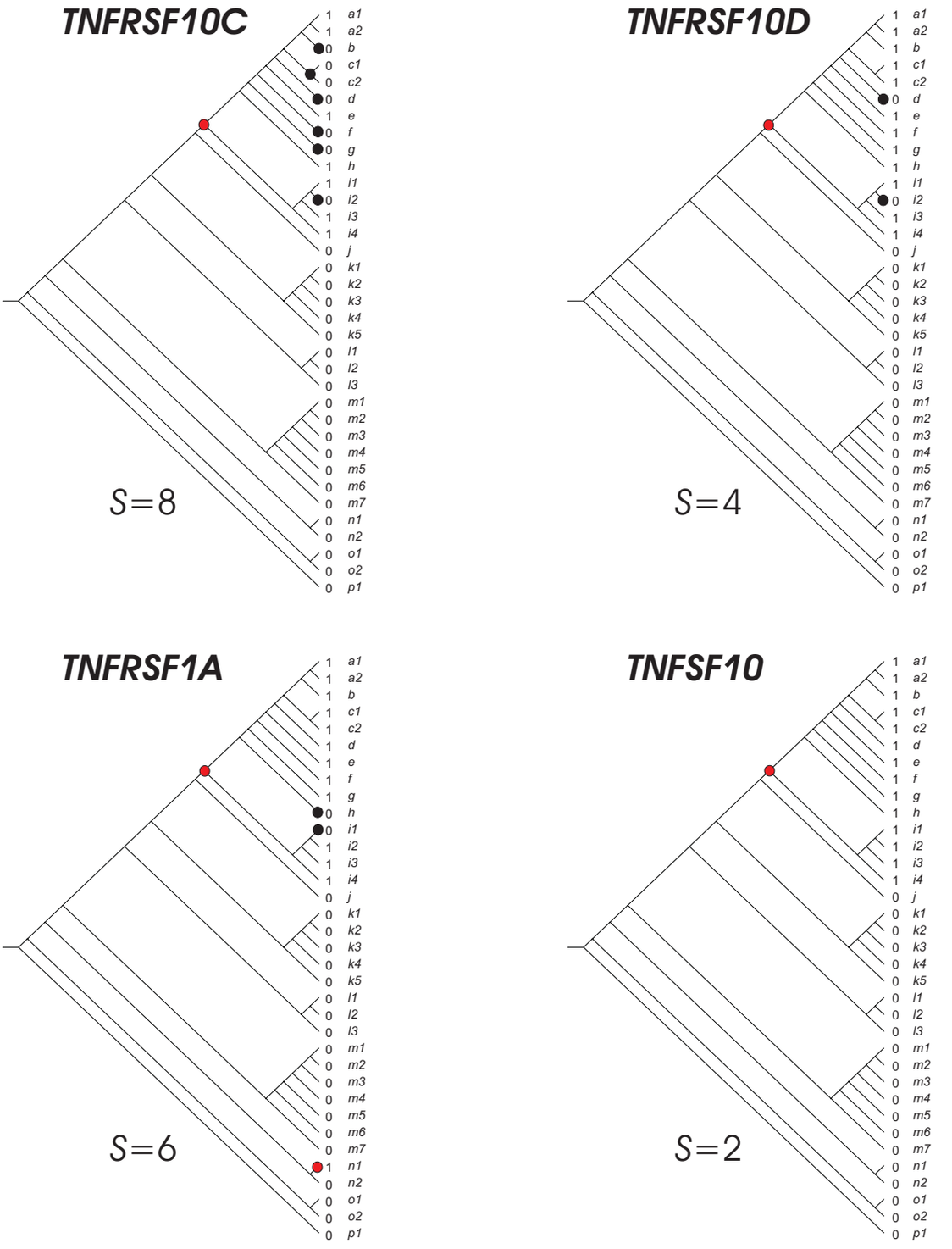
Supplementary Figure S88. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



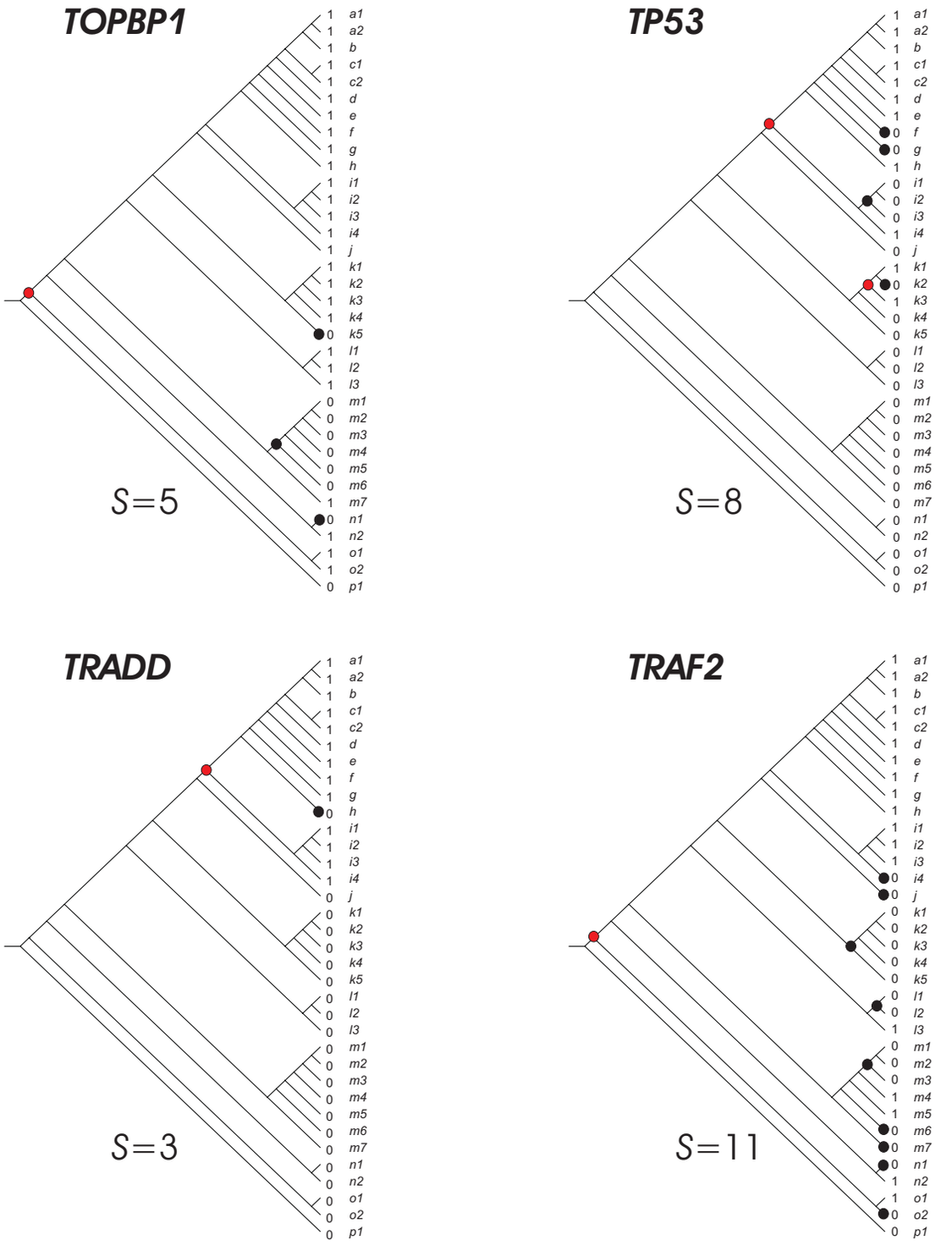
Supplementary Figure S89. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



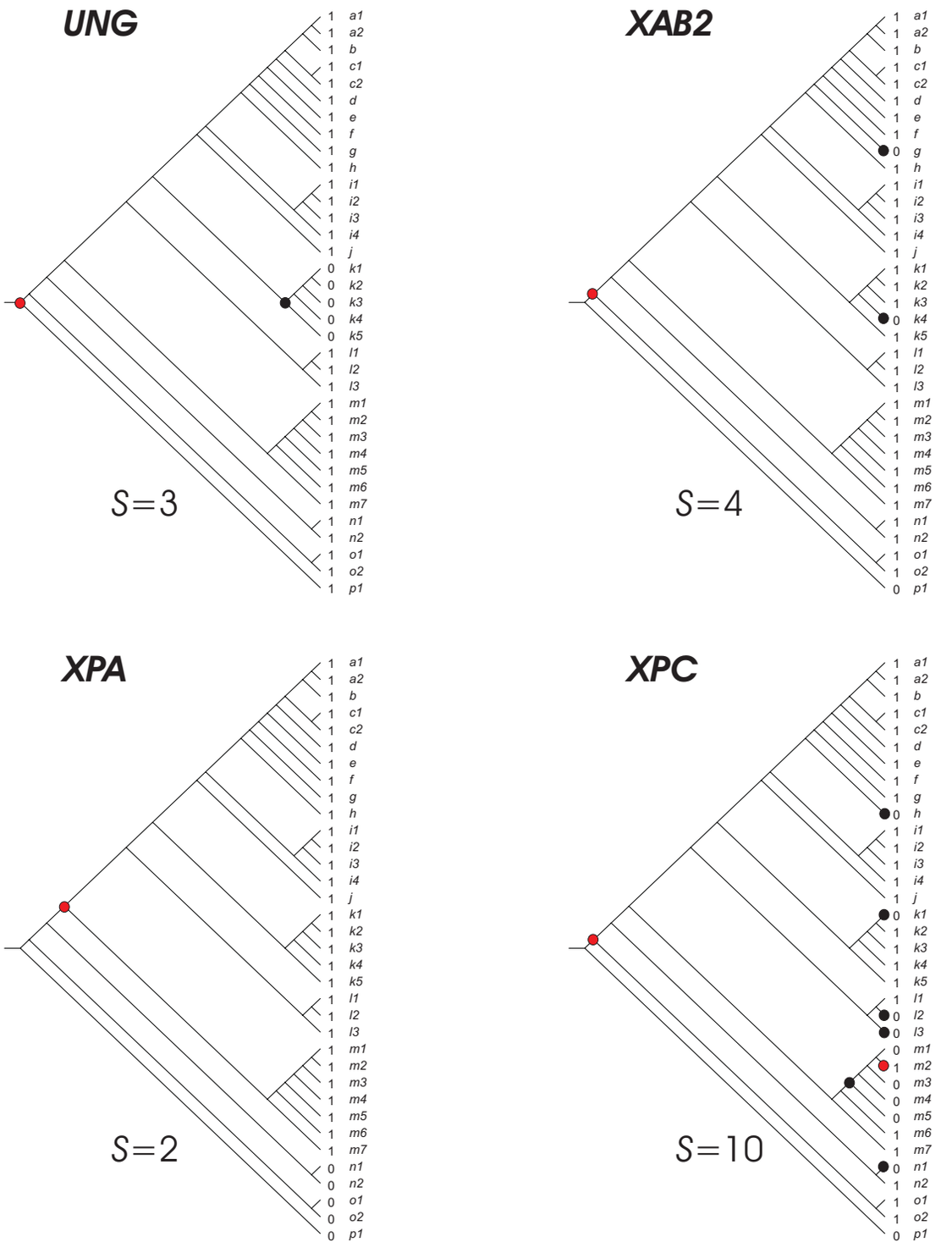
Supplementary Figure S90. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



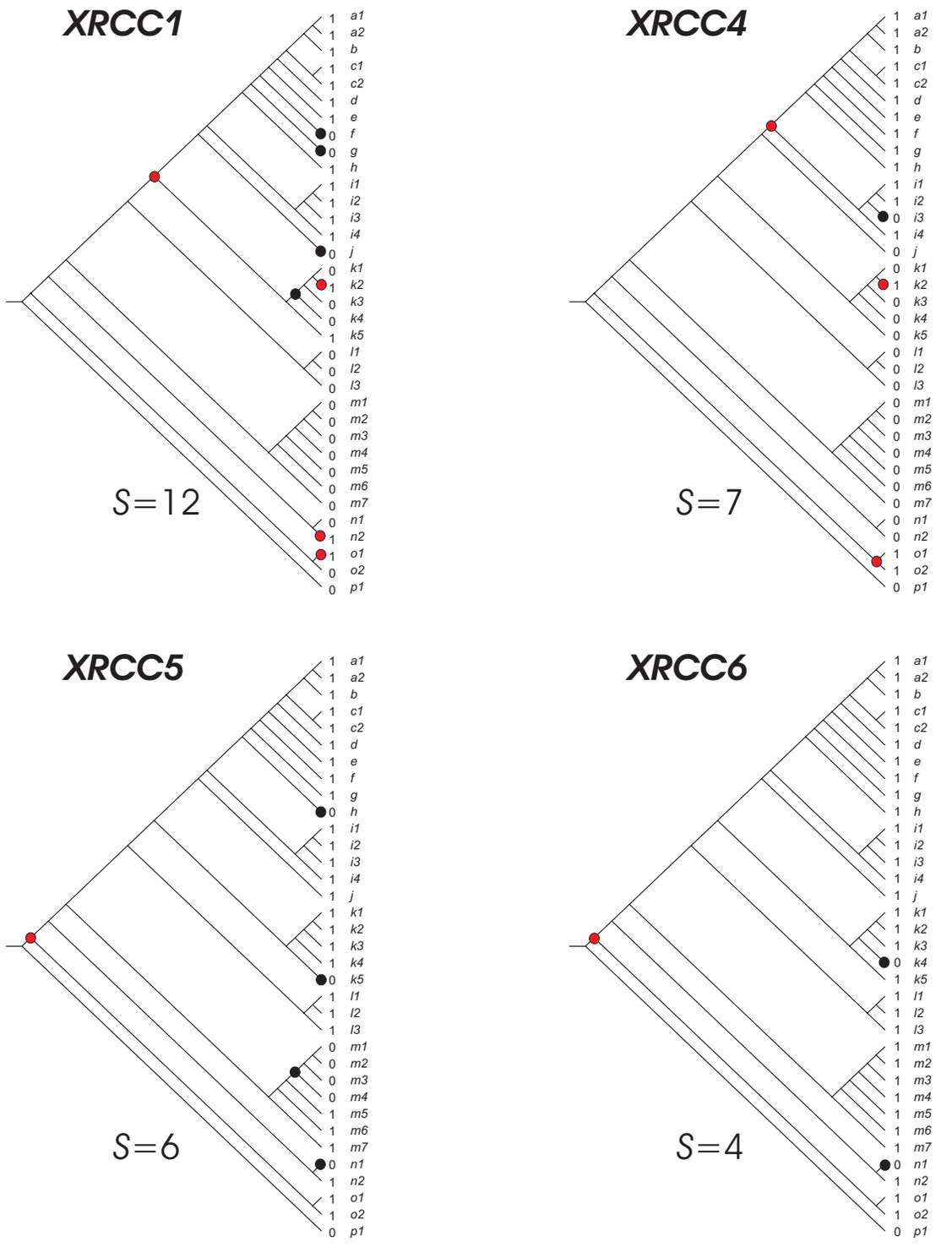
Supplementary Figure S91. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



Supplementary Figure S92. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



Supplementary Figure S93. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.



Supplementary Figure S94. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost unit for gene loss (black nodes) as described in *material and methods*. Branch codes as in Supplementary Figure S50.

2. References

Baldauf, S.L. 2003. The deep roots of eukaryotes. *Science* **300**:1703-1706.

Best, A.A., Morrison, H.G., McArthur, A.G., Sogin, M.L., and Olsen, G.J. 2004. Evolution of eukaryotic transcription: Insights from the genome of Giardia lamblia. *Genome Res.* **14**:1537-1547.

Castro, M.A.A., Mombach, J.C.M., de Almeida, R.M.C., and Moreira, J.C.F. 2007. Impaired expression of NER gene network in sporadic solid tumors. *Nucleic Acids Res.* **35**:1859-1867.

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

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

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284**:2124-2128.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and the Mouse Genome Database Group. 2007. The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.* **35**:D630-D637.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. 2004. A census of human cancer genes. *Nat. Rev. Cancer* **4**:177-183.

Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R. 2003. The Genetic Core of the Universal Ancestor. *Genome Res.* GR-6528.

Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R., et al. 2006. Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.* **34**:D442-D445.

Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., et al. 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414**:450-453

Kunin, V. and Ouzounis, C.A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**:1589-1594.

Letunic, I. and Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**:127-128

Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **33**:4626-4638.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3.**

Pennisi, E. 2003. Drafting a Tree. *Science* **300**:1694

Remm, M., Storm, C.E.V., and Sonnhammer, E.L.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**:1041-1052.

Snel, B., Bork, P., and Huynen, M.A. 2002. Genomes in Flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**:17-25.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**:41.