Universidade Federal do Rio Grande do Sul
Instituto de Ciências Básicas da Saúde
Departamento de Bioquímica
Programa de Pós-graduação em Ciências Biológicas: Bioquímica

DESENVOLVIMENTO DE FERRAMENTAS DE BIOINFORMÁTICA PARA O ESTUDO EVOLUTIVO DE SISTEMAS BIOQUÍMICOS

Doutorando: Rodrigo Juliani Siqueira Dalmolin

Orientadores: Prof. Dr. José Cláudio Fonseca Moreira (Orientador)
Profa. Dra. Rita Maria Cunha de Almeida (Co-orientadora)

Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas: Bioquímica como requisito para a obtenção do grau de Doutor em Ciências Biológicas: Bioquímica.

Porto Alegre, 2012.

Dedico esta Tese aos meus pais Mauro e Ani, que infelizmente partiram ao longo da minha formação; à minha esposa Joana, que é uma das grandes responsáveis por esta realização; e ao meu filho Francisco, que em breve chegará.

“Nature can produce complex structures even in simple situations, and can obey simple laws even in complex situations.”

Goldenfeld and Kadanoff. Science 284, 87 (1999)

Agradecimentos

Ao Glorioso Professor **Zé Cláudio**, pelo papel fundamental que exerceu ao longo de toda a minha formação, por ter me acolhido em seu laboratório em duas oportunidades, pela liberdade e confiança durante a execução do trabalho, pela coragem em orientar uma tese relativamente distante de suas linhas de pesquisa originais e, acima de tudo, pela amizade cultivada ao longo dos últimos 12 anos.

À Professora **Rita**, pela co-orientação, pela confiança, pela inestimável contribuição intelectual e científica e pela amizade.

Ao colega **Mauro Castro**, pelos valiosos ensinamentos, pela contribuição indispensável no desenvolvimento desta tese, pelo coleguismo e pela amizade.

Ao colega **Zé Luiz Rybarczyk Filho**, pelo papel fundamental no desenvolvimento das ferramentas de bioinformática aplicadas nesta tese e pela amizade.

Aos colegas **Luís Henrique Souza** e **Ricardo Albanus**, pela pronta colaboração.

Ao Professor **Daniel Gelain**, por me aproximar do Departamento de Bioquímica.

A todos os colegas do CEEO, pela agradável e divertida convivência.

A toda a comunidade do Departamento de Bioquímica.

A todos os colegas do Instituto de Física.

À UFRGS e ao PPG em Ciências Biológicas: Bioquímica, pela excelente formação.

Às agências de Fomento, CAPES e CNPq, e a população Brasileira, pelo financiamento da minha formação.

Muito Obrigado!

Índice

Parte I

Resumo

O crescente corpo de informações gerado pelo desenvolvimento de técnicas de alto-desempenho, como sequenciamento de DNA em larga escala, técnicas de microarranjo de DNA, hibridização de proteínas, etc., tem evidenciado uma intrincada relação entre os diversos personagens que compõe os sistemas biológicos. Alguns dos sistemas bioquímicos presentes em organismos modernos surgiram há bilhões de anos e estavam presentes em organismos primitivos, ao passo que determinados sistemas são mais recentes e específicos de alguns grupos taxonômicos. O entendimento das relações entre os diferentes personagens dos sistemas biológicos apresenta-se como fundamental para a compreensão da vida e a avaliação dos aspectos evolutivos que permearam a constituição dos sistemas bioquímicos e suas intrincadas inter-relações pode auxiliar sobremaneira no estudo da biologia. Diversas teorias encontram-se bem estabelecidas no estudo evolutivo em nível de espécies e populações. Da mesma maneira, há um extenso acervo bibliográfico acerca da evolução de genes individuais. Entretanto, o surgimento, estabelecimento e evolução dos sistemas bioquímicos permanecem escassamente estudados. Na presente tese, partimos da análise de dois sistemas bioquímicos, o sistema de apoptose e o sistema de estabilidade genômica, os quais são bastante associados em mamíferos. Apesar da íntima relação entre esses sistemas, eles foram originados em momentos diferentes da evolução. Buscamos reconstruir o cenário evolutivo que uniu os sistemas de apoptose e estabilidade genômica, onde encontramos uma relação direta entre ancestralidade, essencialidade e clusterização. Os resultados também sugerem uma relação inversa entre essas três características e plasticidade. A análise de plasticidade efetuada na rede de apoptose e estabilidade genômica foi ampliada para 4850 famílias de proteínas em 55 eucariotos, apresentando basicamente os mesmos resultados, indicando um mecanismo geral de evolução do genoma. Subsequentemente, propusemos um modelo matemático de crescimento do genoma onde a novidade genética surge por duplicação de genes muito conectados e pouco clusterizados. A rede artificial obtida mimetiza diversos aspectos topológicos das redes biológicas conhecidas. Os resultados analisados em conjunto sugerem um mecanismo geral de evolução do genoma, onde a novidade genética surge na porção mais plástica do genoma, basicamente por duplicação gênica. Essa duplicação ocorre prioritariamente nos *hubs* intermodulares.

Abstract

The increasing body of information generated by high-throughput techniques, such as DNA sequencing, genome-wide microarray, and two-hybrid system, has unveiled an intricate relationship among different components of biological systems. Some of the biological systems found in modern organisms have their origins billion years ago and were present in primitive organisms. On the other hand, some biological systems are more recent and specifically related to some taxa. The characterization of the relationships involving the different components of biological systems is crucial to the understanding of life. Additionally, the evaluation of evolutionary aspects which work in biochemical systems construction, modeling their intricate relationship, could help improve biological research field. Several theories are well-established in evolutionary research of species and population. Likewise, there is plenty of bibliography concerning individual gene evolution. However, there is paucity of data concerning the origin, establishment, and evolution of entire biological systems. In the present thesis, we start by analyzing two biochemistry systems: apoptosis and genome stability. These systems are considerably associated in mammals. Despite its entangled functioning, each system has emerged in different points of evolution. We reconstructed the evolutionary scenario which entangled both systems. We found a direct relationship among ancestrality, essentiality, and clustering. Our results also suggest an inverse relationship of these three proprieties with plasticity. The same plasticity analysis used in apoptosis and genome stability systems was amplified to 4850 gene families in 55 eukaryotes, showing basically the same results. It suggests a general mechanism of genome evolution. We then propose a genome growth model where genetic novelty arrives through gene duplication of highly connected but not so clustered genes. The resulting artificial network reproduces several known topological aspects of biological networks. The results, when simultaneously analyzed, suggest general genome evolution mechanisms, where the genetic novelty arrives in more plastic area of the genome, basically by gene duplication. That duplication occurs mainly in intermodular hubs.

Lista de Abreviaturas

BER – Reparo por Excisão de Base (do inglês, *Base Excision Repair*)

CC – Coeficiente de Clusterização

COG – Cluster de Grupos Ortólogos (do inglês, *Cluster of Orthologous Groups*)

EPI – Índice de Plasticidade Evolutiva (do inglês, *Evolutionary Plasticity Index*)

K – Conectividade ou grau

KEGG – Enciclopédia de Genes e Genomas de Kyoto (do inglês, *Kyoto Encyclopedia of Genes and Genomes*)

KOG – Cluster Eucariótico de Grupos Ortólogos (do inglês, *Eukaryotic Cluster of Orthologous Groups*)

NCBI – National Center for Biotechnology Information

NER – Sistema de Excisão de Nucleotídeos (do inglês, *Nucleotide Excision Repair*)

OG – Grupo de ortólogos (do inglês, *Orthologous Groups*)

Introdução

O Estudo dos Sistemas Biológicos

Existem diversos níveis de organização dos sistemas biológicos: desde sistemas bioquímicos com funcionamento restrito a organelas específicas, até células, organismos pluricelulares e comunidades biológicas envolvendo um grande número de indivíduos de diferentes espécies. Embora cada uma dessas unidades organizacionais represente um diferente grau hierárquico da vida, é difícil dizer quais desses citados níveis apresentam uma constituição mais ou menos complexa. De fato, os sistemas biológicos representam o caso mais extremo de complexidade do universo conhecido e a sua compreensão constitui-se em um desafio de equivalente magnitude (Goldenfeld and Kadanoff, 1999).

Historicamente, a estratégia de estudo da biologia tem consistido na segmentação de cada sistema de modo a compreendê-lo individualmente. Dessa forma, diferentes níveis organizacionais têm sido estudados por diferentes ramos das ciências biológicas, como ecologia, fisiologia, bioquímica, biologia molecular, etc. Da mesma maneira, sistemas celulares têm sido desmembrados, onde cada parte fundamental (*i.e.* biomoléculas) é estudada isoladamente a fim de se compreender o funcionamento do sistema como um todo. Essa estratégia reducionista tem sido utilizada durante décadas e trouxe inegáveis avanços na compreensão da biologia. Entretanto, uma das características de sistemas complexos é a presença de propriedades emergentes, onde o todo não pode ser explicado simplesmente pela soma de suas partes (Amaral and Ottino, 2004). Dessa forma, metodologias que procurem compreender os sistemas como uma unidade, avaliando seus componentes em conjunto, podem contribuir para o entendimento da biologia.

A partir da década passada, a *biologia de sistemas* tem ganhado força como um promissor ramo das ciências biológicas que tem se dedicado a entender os sistemas a partir do estudo das interações entre seus componentes (Pujol et al., 2010). O seu desenvolvimento tem sido favorecido devido, principalmente, ao desenvolvimento de técnicas de alto desempenho, capazes de gerar uma grande quantidade de dados em um curto espaço de tempo. A crescente quantidade de genomas completos disponíveis, aliados a técnicas como microarranjo de DNA e hibridização de proteínas, por exemplo, proporcionam uma avaliação em larga escala do *status* fisiológico de células e tecidos (Barabasi and Oltvai, 2004; Rybarczyk-Filho et al., 2010). Essa crescente quantidade de dados gerada proporcionou a aplicação da teoria de grafos sobre dados biológicos a partir da representação dos sistemas biológicos como redes de interações (Yamada and Bork, 2009).

A relação entre os diversos personagens de sistemas biológicos tem sido historicamente descrita de diferentes formas. Em bioquímica, o conceito de *rotas* é o mais comumente utilizado. Em uma rota bioquímica, a descrição das relações entre os diferentes agentes (*e.g.* substratos, enzimas, etc.) é feita a partir de um fluxo de massa, onde geralmente os substratos são sequencialmente modificados por enzimas, com a entrada e a saída de subprodutos (Figura 1A). A representação de sistemas bioquímicos como redes de interação tem surgido como uma forma alternativa de avaliar e entender as relações entre biomoléculas. Dentre as diferentes redes de interações, estão aquelas que envolvem diferentes classes de moléculas, como por exemplo, enzimas, substratos e cofatores (chamadas de redes metabólicas – Figura 1B), e aquelas envolvendo somente interações entre proteínas (chamadas de redes de interação proteína-proteína – Figura 1C). Ao passo que em uma rota bioquímica a direção das reações é uma informação

fundamental, em uma rede de interações as características topológicas assumem o protagonismo.

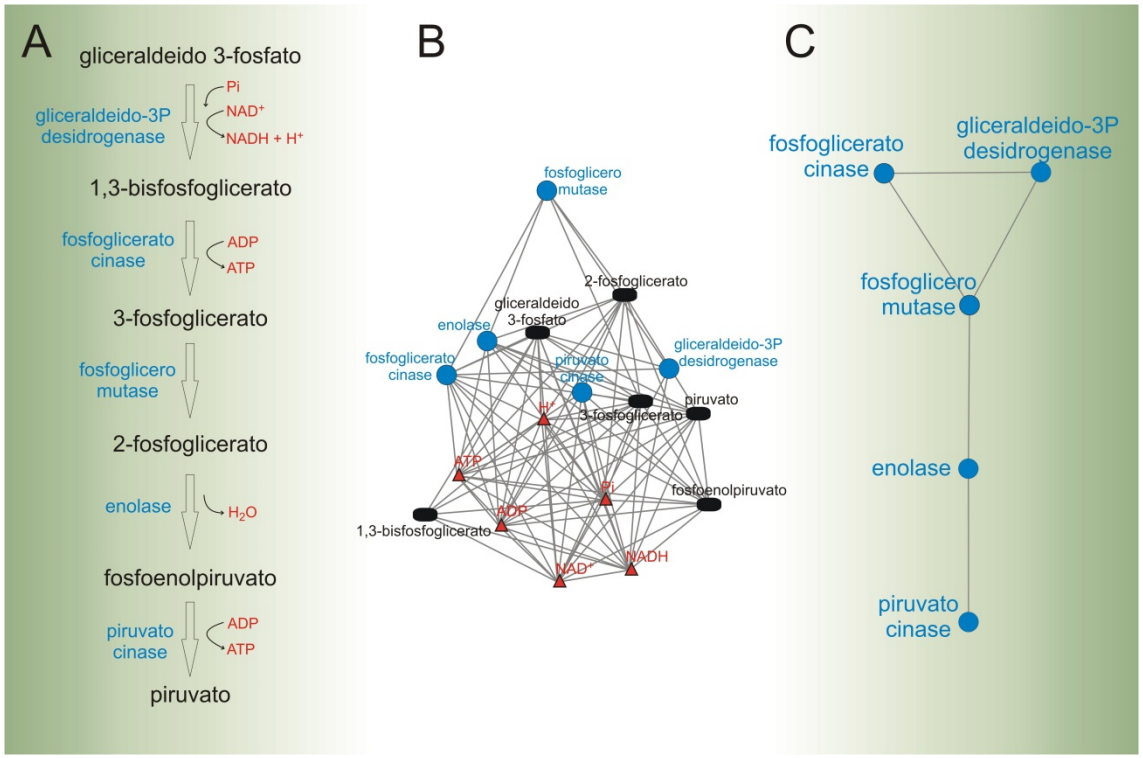


Figura 1. Três representações distintas da segunda fase da glicólise (fase de retorno). Rota bioquímica clássica (A), rede metabólica (B) e rede de interações proteína-proteína (C). As setas representam o fluxo de massa das modificações sofridas pelas moléculas e as arestas representam interações entre as moléculas.

Em uma rede de interações proteína-proteína (ou rede de interação proteica) os nós representam proteínas e as arestas (ou *links*) representam a presença de interação entre os nós da rede (Figura 1C). Uma variação da rede de interação proteica é a rede de interação gênica, onde os nós representam genes e as arestas representam a interação entre seus produtos proteicos. Essa interação pode ser de diferentes naturezas como, por exemplo, a fosforilação direta de uma proteína pela outra, a ligação física entre duas proteínas ou a participação de ambas em uma mesma rota bioquímica (Harrington et al., 2008). Consequentemente, a informação acerca da presença ou ausência de ligação entre um par de determinadas proteínas advém de diferentes fontes e essa grande quantidade de informação tem sido organizada e disponibilizada em diferentes bancos

de dados. De fato, o aumento da geração de dados sobre sistemas biológicos foi acompanhado do surgimento de repositórios destinados a organizar e disponibilizar a informação biológica.

Dentre os principais repositórios sobre interação proteica encontra-se o banco de dados STRING (http://string-db.org/) o qual integra a informação de diferentes repositórios reunindo uma vasta informação acerca de interações proteicas (Szklarczyk et al., 2011). O repositório STRING possui informação sobre interação proteica de mais de mil espécies e disponibiliza os dados de uma forma organizada, de modo que o usuário pode definir qual tipo de interação lhe é de interesse (participação em uma mesma rota, interação física, co-ocorrência, coexpressão, etc.), bem como o grau de confiança das ligações inferidas. Além de repositórios contendo informações acerca de interações proteicas, existe um sortimento de bancos de dados contendo informações sobre os mais diversos processos e biomoléculas, que são de grande utilidade no estudo de sistemas biológicos. Dentre esses, o repositório KEGG (Enciclopédia de Genes e Genomas de Kyoto, do inglês, *Kyoto Encyclopedia of Genes and Genomes* - http://www.genome.jp/kegg/) merece destaque aqui devido a sua grande utilidade no estudo de sistemas bioquímicos, bem como sua extensa utilização nesta tese. Esse banco de dados possui um vasto repertório de informações, como nomenclatura de enzimas, metabólitos, dados sobre alterações bioquímicas em estados patológicos, além da descrição detalhada de cerca de 400 rotas de referência e suas particularidades em centenas de organismos (Ogata et al., 1999).

Diversas medidas têm sido propostas para compreender os padrões dos sistemas biológicos, baseados principalmente na topologia das redes de interações. Avaliação de propriedades dos nós (como a conectividade ou grau), bem como propriedades da rede (como coeficiente de clusterização e centralidade) têm sido largamente utilizadas no

estudo dos sistemas biológicos. A conectividade representa o número de nós da rede com o qual um determinado nó interage (Yamada and Bork, 2009). A conectividade de uma proteína, do ponto de vista bioquímico, pode ser entendida como o número de outras proteínas com a qual ela interage. Algumas proteínas participam de processos bioquímicos específicos e interagem com um número limitado de proteínas. Como exemplo, temos a enzima CCS, uma chaperona evolvida especificamente na entrega de cobre para a enzima antioxidante SOD1(Culotta et al., 1997). Portanto, essa proteína apresenta apenas uma ligação em redes biológicas (Gelain et al., 2009). Como um exemplo oposto, temos o caso da proteína p53, a qual participa de diferentes rotas bioquímicas (Sengupta and Harris, 2005). Consequentemente, essa proteína apresenta alta conectividade em sistemas biológicos (Castro et al., 2007). O coeficiente de clusterização indica se os nós com os quais um determinado nó interage são também conectados entre si ou não. Por exemplo, proteínas que fazem parte de complexos proteicos apresentam coeficiente de clusterização alto, visto que todos os membros do complexo interagem entre si, caracterizando um módulo biológico. De modo contrário, proteínas pleiotrópicas que participam de diversos processos metabólicos, apresentam baixo coeficiente de clusterização por interagirem com várias outras proteínas as quais não interagem mutuamente entre si (Yamada and Bork, 2009).

Diversos padrões de redes artificiais têm sido propostos buscando modelos que expliquem tanto a organização das redes biológicas como os processos que regem a evolução dos genomas (Barabasi and Oltvai, 2004; Yamada and Bork, 2009). O mais notável estudo acerca de redes biológicas foi desenvolvido por Barabási e colaboradores, onde a partir de dados de interação entre genes/proteínas, foi proposto um modelo de rede livre de escala como arquitetura das redes biológicas (Barabasi and Albert, 1999). No mesmo trabalho, os autores sugeriram que a estrutura observada das

redes biológicas se deve ao crescimento do genoma a partir da ligação de novos genes a genes preexistentes na rede. Quanto maior a conectividade de um gene, maior a probabilidade de um novo gene se ligar a ele. Entretanto, os mecanismos geradores dos novos genes a serem agregados à rede não foram abordados na ocasião. Mais de uma década separa as descrições de Barabási dos dias de hoje e a quantidade de dados gerados nestes últimos anos aumentou substancialmente e continua avançando. Um dos maiores desafios da ciência na era pós-genômica consiste justamente em depurar e compreender a grande quantidade de dados gerados, a fim de transformá-los em informação.

O Estudo evolutivo dos Sistemas Bioquímicos

Desde os primeiros ensaios acerca da biologia evolutiva, os estudos têm se concentrado nas características fenotípicas das espécies. Embora, do ponto de vista prático, a seleção natural ocorra no nível fenotípico, são as alterações genotípicas que de fato são mantidas ou descartadas pelo processo evolutivo. Levando em conta que a dinâmica dos processos evolutivos deixa marcas nos genomas, o material genético das espécies pode ser, portanto, considerado um registro da história evolutiva. Embora esse registro seja incompleto e fragmentado, a utilização de técnicas de biologia molecular representou um avanço no estudo evolutivo, outrora baseado quase que estritamente em registros anatômicos, morfológicos e paleontológicos (Dan Graur and Wen-Hsiung Li, 2000).

O principal foco de estudo da evolução molecular compreende as alterações sofridas na sequência gênica. Diversas são as alterações que podem modificar a

informação genética de uma espécie: desde mutações pontuais até duplicações ou deleções completas de grandes porções do genoma. Tais alterações podem ou não ser selecionadas, dependendo do impacto que causarem no sistema do qual fazem parte (Koonin and Wolf, 2010). Além disso, alterações genéticas neutras em relação à adaptabilidade do organismo que as carrega podem ser mantidas através de processos randômicos de fixação (Kimura, 1991).

As primeiras pesquisas no campo da evolução molecular, em meados da década de 60, eram focadas basicamente em uma abordagem mendeliana. Dessa forma, a maioria dos estudos investigava alterações entre genes homólogos dentro da mesma espécie (Nei, 2005). Com o passar dos anos, os estudos foram expandidos para espécies próximas, culminando, mais recentemente, na avaliação de grandes quantidades de genes ou proteínas em diferentes espécies. Estes estudos se fizeram possíveis, principalmente, graças à grande quantidade de dados gerados por técnicas de alto-desempenho, sobremaneira as técnicas de sequenciamento que permitiram a determinação do genoma completo de um grande número de espécies.

A análise de uma grande quantidade de genomas sequenciados incrementou a discussão acerca das relações evolutivas entre proteínas/genes completos de diferentes espécies. Em conjunto com o avanço dos estudos direcionados ao entendimento destas relações, uma diversa nomenclatura povoou os trabalhos envolvidos no estudo de famílias de genes. Termos como parálogos, inparálogos, outparálogos, ortólogos, co-ortólogos, pseudo-ortólogos, etc., não são raros em artigos sobre esse tema (Koonin, 2005). Cabe aqui uma pequena explanação sobre os conceitos de ortologia empregados nesta tese. Genes ortólogos são aqueles herdados verticalmente durante o processo de especiação e genes parálogos são aqueles originados a partir de um episódio de duplicação gênica (Koonin, 2005). Genes ortólogos entre si são comumente associados

a funções semelhantes (Chen and Jeong, 2000). Entretanto, a associação direta entre ortologia e funcionalidade é controversa, já que o tempo de divergência entre genes ortólogos pode influenciar na sua similaridade funcional (Studer and Robinson-Rechavi, 2009).

A figura 2 mostra uma árvore filogenética hipotética, onde temos três espécies (espécie a, espécie b e espécie c). Durante o processo de especiação, as três espécies herdaram o gene *X* da espécie ancestral. Tais genes (X_2, X_3, X_4 *e* X_5) são considerados ortólogos em relação ao gene X_1. Entretanto, durante o processo de especiação que deu origem a espécie c, ocorreu a fixação de um episódio de duplicação gênica. Desta forma, a espécie c apresenta dois genes homólogos, X_4 e X_5, que são considerados parálogos entre si. Da mesma maneira, tanto o gene X_4 quanto o gene X_5 são considerados ortólogos em relação aos genes X_3 e X_2.

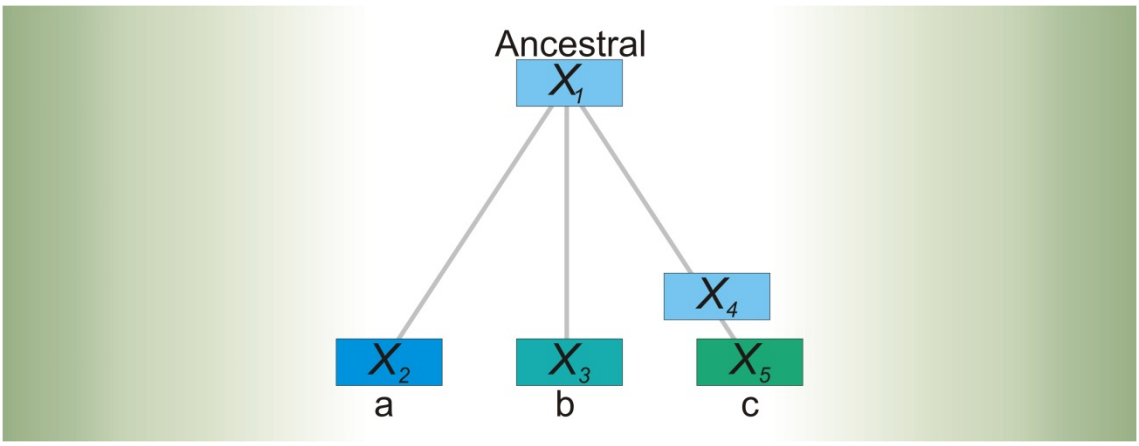


Figura 2. Representação de uma árvore filogenética hipotética. Cada vértice da árvore representa uma espécie e os retângulos representam genes. As diferenças entre os genes, fruto do processo de especiação, é representada pelas diferentes cores. Adaptado de Koonin, 2005.

Acompanhando o crescente corpo de genomas sequenciados, surgiram diversos repositórios objetivando organizar os dados de ortologia, cada um se valendo de diferentes algoritmos no intuito de identificar relações evolutivas entre genes de

diferentes espécies (Kanehisa and Goto, 2000; Li et al., 2003; O'Brien et al., 2005; Tatusov et al., 2001). Dentre eles, o mais relevante para a presente tese é o projeto *Cluster of Orthologous Groups* (COG) (http://www.ncbi.nlm.nih.gov/COG/), desenvolvido pelo *National Center for Biotechnology Information* (NCBI) (Tatusov et al., 1997) e ampliado pelo repositório STRING (http://string-db.org/). Atualmente o repositório STRING utiliza 1133 organismos totalmente sequenciados para a construção dos clusters de grupos de ortólogos, além de apresentar os clusters de grupos de ortólogos de eucariotos (KOG – do inglês, *Eukaryotic Clusters of Orthologous Group*) (Szklarczyk et al., 2011). Basicamente, o banco COG utiliza triangulações entre proteínas de diferentes organismos para a formação dos clusters. Ele parte do princípio que, após a identificação de parálogos óbvios, se três proteínas de espécies diferentes são mais similares entre si do que com qualquer outra dentro do próprio genoma, essas proteínas formam um grupo de ortólogos (Koonin, 2005). Para a formação de um grupo de ortólogos são necessárias ao menos três proteínas com similaridade recíproca. Entretanto, um grupo de ortólogos pode congregar centenas de proteínas as quais são, teoricamente, descendentes de uma mesma proteína ancestral.

A partir dos conceitos de ortologia, é possível traçar um panorama da história evolutiva de sistemas bioquímicos inteiros. A despeito de haver questionamentos em relação a imputar correlações funcionais a um grupo de proteínas a partir da sua similaridade estrutural, valiosas informações podem ser obtidas a partir do padrão de distribuição de proteínas ortólogas em diferentes espécies. Observemos um exemplo hipotético onde um grupo de proteínas apresenta o padrão de herança representado na figura 3. De acordo com o exemplo, a *espécie 1* apresenta um conjunto de proteínas (A, B, C, D e E), as quais sabidamente atuam conjuntamente em um mesmo sistema bioquímico nesta espécie. O fato de tanto a *espécie 2* quanto a *espécie 4* apresentarem

um conjunto de genes ortólogos em relação aos genes codificantes das proteínas da *espécie 1*, indica um padrão de coherança deste grupo de proteínas. Além disso, a *espécie 3* não apresenta ortólogos de nenhuma destas proteínas. Tais padrões de herança, onde conjuntos de proteínas são simultaneamente presentes ou ausentes em diferentes espécies, sugere que as mesmas são funcionalmente relacionadas (Glazko and Mushegian, 2004).

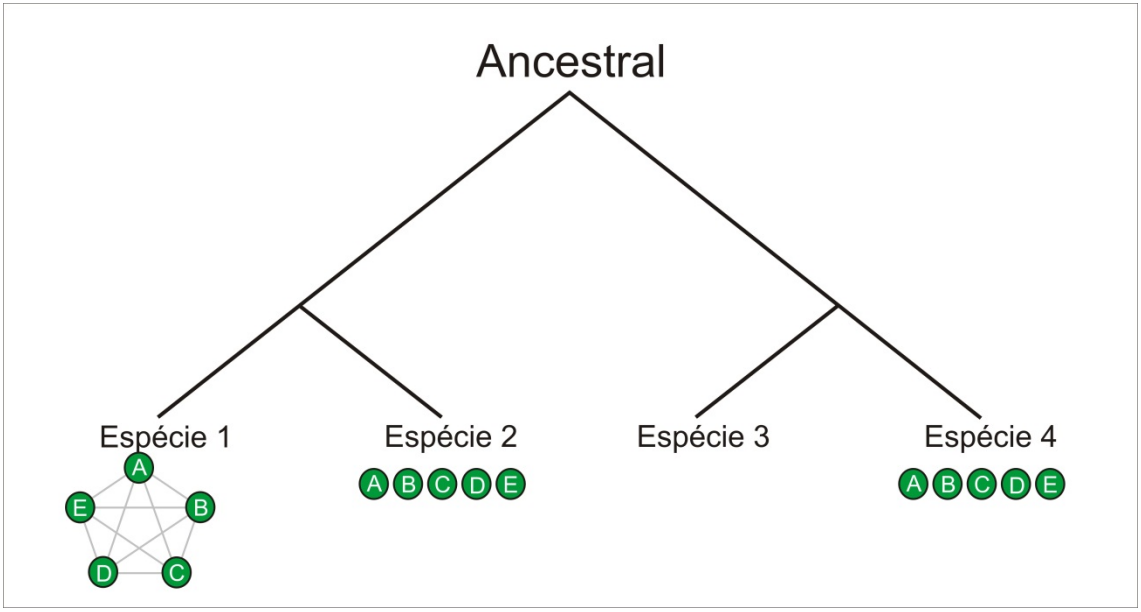


Figura 3. Representação de uma árvore filogenética hipotética, representando a herança em conjunto de um grupo de proteínas. Cada vértice da árvore representa uma espécie e cada círculo representa uma proteína. As arestas, em cinza, indicam que há interação entre as proteínas.

Surgimento de novidade genética e crescimento do genoma.

De acordo com Ernst Mayr, a seleção natural assemelha-se mais a um processo de descarte, onde os menos aptos são eliminados da população, do que a um processo de seleção propriamente dito, onde somente os mais aptos são selecionados (Mayr, 2005). Dessa forma, alterações as quais diminuam significativamente a adaptabilidade do indivíduo são fortemente constrangidas, ao passo que tanto modificações neutras quanto modificações que aumentem a adaptabilidade podem ser fixadas; estas últimas com maior probabilidade. Em um sistema bioquímico complexo, onde os diferentes pares

atuam em conjunto na execução de tarefas metabólicas, uma alteração aleatória em um dos seus componentes apresenta maior probabilidade de ser deletéria do que de aumentar a adaptabilidade do sistema. Por conseguinte, outros mecanismos além de modificações em genes funcionais são fundamentais para a evolução dos sistemas bioquímicos. Atualmente sabe-se que diversos mecanismos moleculares, como transferência horizontal, modificações em regiões não codificantes, elementos transponíveis, duplicação gênica, etc., contribuem para o surgimento de novidade genética. Dentre eles, processos de duplicação gênica são reconhecidamente os mais importantes (Long and Thornton, 2001; Long et al., 2003).

A ideia de que eventos de duplicação gênica proporcionam a principal fonte de matéria-prima para o surgimento de novos genes ganhou força a partir da década de 70, com a publicação do famoso trabalho de Susumu Ohno intitulado *Evolution by gene duplication* (Kaessmann, 2010). Uma vez duplicado, a pressão seletiva é diminuída em pelo menos uma das cópias do gene, já que a função original deste é mantida pela outra cópia. Enquanto uma cópia é mantida funcional, a outra pode sofrer mutações que podem levá-la a desenvolver novas funções. Os episódios de duplicação acontecem randomicamente no genoma; entretanto, não necessariamente a duplicação será fixada. Muitas vezes a nova cópia é alvejada por mutações, podendo tornar-se um pseudogene ou perder completamente a identidade estrutural com o gene a qual teve origem. Diversas teorias procuram explicar a fixação dos genes formados a partir de duplicação. Quase que invariavelmente elas associam a fixação da duplicação ao desenvolvimento de uma nova função por uma das cópias (Innan and Kondrashov, 2010).

O surgimento de uma nova função a partir de um episódio de duplicação pode ser favorecido quanto maior for a plasticidade funcional do produto do gene em questão. Uma proteína que exerce mais de uma função em diferentes sistemas bioquímicos pode

apresentar um “conflito adaptativo”, não se especializando em nenhuma das funções executadas. Ambas as funções podem ser beneficiadas por uma duplicação, onde cada cópia pode especializar-se em uma das funções outrora executadas por uma única proteína (Deng et al., 2010). Além disso, é razoável supor que proteínas que apresentem motivos funcionais que podem servir a mais de um processo metabólico - mesmo que não o façam previamente - teriam uma maior probabilidade de fixar um eventual episódio de duplicação. Em contrapartida, duplicações que ocorram em genes envolvidos em sistemas de baixa plasticidade, pouco tolerantes a alterações, podem não serem fixadas ou até mesmo serem constrangidas. De uma forma geral, é presumível que o surgimento de novidade genética aconteça em sistemas de maior plasticidade, mais tolerantes a mudanças, do que em sistemas de baixa plasticidade, onde uma alteração tem grande probabilidade de ser deletéria.

Os mecanismos geradores de novidade genética são cruciais para o crescimento dos genomas e o entendimento de quais são e onde atuam estes mecanismos é fundamental para o entendimento da dinâmica do processo evolutivo. Portanto, é aconselhável que um modelo evolutivo que se proponha a explicar a evolução do genoma leve em conta os mecanismos moleculares envolvidos no surgimento de novidade genética. Como mencionado anteriormente, Barabási e colaboradores propuseram um modelo de crescimento do genoma o qual levava em conta o grau de um gene presente na rede para definir a probabilidade de um novo gene conectar-se ou não a ele. Apesar da fundamental importância deste trabalho como pioneiro na análise das propriedades de redes biológicas, tal modelo de crescimento não leva em conta a origem dos novos genes (Barabasi and Albert, 1999), mas simplesmente uma das propriedades dos genes presentes na rede.

Mais de 10 anos passados do modelo proposto por Barabási e Albert, Vázquez e colaboradores propuseram um modelo baseado na hipótese de que o crescimento do genoma se dá por um processo de duplicação seguido de divergência, o que vem ao encontro das teorias evolutivas em relação ao surgimento de novidade genética (Vázquez et al., 2003). Este modelo, chamado de Duplicação-Divergência, parte do princípio de que as redes biológicas são livres de escala e utiliza um algoritmo onde os genes são aleatoriamente escolhidos e duplicados. Uma particularidade do modelo consiste no fato de que o novo nó da rede (o qual representa um gene duplicado) herda as mesmas ligações que o nó parental possuía. Após a duplicação, existe uma probabilidade de mutação, onde os nós podem perder algumas das ligações herdadas. O modelo Duplicação-Divergência representou um avanço no sentido de incluir o mais relevante mecanismo conhecido envolvido no surgimento de novidade genética (*i.e.* duplicação gênica) em um algoritmo de crescimento de redes. A crítica que pode ser feita ao modelo Duplicação-Divergência advém do fato de os nós serem randomicamente escolhidos para duplicar, já que sabidamente a fixação de uma duplicação gênica não é randômica e dependerá das características funcionais do gene duplicado (Conant and Wolfe, 2008). Entretanto, não há consenso acerca de quais são as exatas características de um gene que aumentariam a probabilidade de fixação de uma duplicação. E a conversão destas características em propriedades de rede não necessariamente é uma tarefa simples.

Objetivos do Trabalho

Objetivo geral

Dado que o entendimento dos processos evolutivos que desenharam os sistemas bioquímicos atuais pode contribuir sobremaneira na compreensão do funcionamento dos mesmos, a presente tese tem como objetivo investigar as relações que regem o surgimento e a evolução de sistemas bioquímicos, propondo modelos e ferramentas de bioinformática para auxiliar nesta investigação.

Objetivos específicos

1- Identificar a partir de uma rede conhecida a origem evolutiva das redes humanas de apoptose e estabilidade genômica;
2- Propor uma medida de plasticidade e conservabilidade evolutiva;
3- Propor um mecanismo de evolução do genoma que explique a dinâmica observada nos processos evolutivos.

Parte II

Capítulo I

Evolutionary origins of human apoptosis and genome-stability gene networks.

Artigo científico publicado no periódico *Nucleic Acids Research* (doi: 10.1093/nar/gkn636).

Published online 2 October 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 19* **6269–6283**
doi:10.1093/nar/gkn636

Evolutionary origins of human apoptosis and genome-stability gene networks

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

[1]Bioinformatics Unit, Department of Biochemistry, Federal University of Rio Grande do Sul (UFRGS), Rua Ramiro Barcelos 2600-anexo, Porto Alegre 90035-003, [2]Department of Biological Sciences, Lutheran University of Brazil, Gravataí 94170-240, [3]Department of Physics, Federal University of Santa Maria (UFSM), Santa Maria 97105-900 and [4]Institute of Physics, Federal University of Rio Grande do Sul (UFRGS), Avenida Bento Gonçalves 9500, Porto Alegre 91501-970, Caixa Postal 15051, Brazil

Received May 23, 2008; Revised September 14, 2008; Accepted September 15, 2008

ABSTRACT

Apoptosis is essential for complex multicellular organisms and its failure is associated with genome instability and cancer. Interactions between apoptosis and genome-maintenance mechanisms have been extensively documented and include transactivation-independent and -dependent functions, in which the tumor-suppressor protein p53 works as a 'molecular node' in the DNA-damage response. Although apoptosis and genome stability have been identified as ancient pathways in eukaryote phylogeny, the biological evolution underlying the emergence of an integrated system remains largely unknown. Here, using computational methods, we reconstruct the evolutionary scenario that linked apoptosis with genome stability pathways in a functional human gene/protein association network. We found that the entanglement of DNA repair, chromosome stability and apoptosis gene networks appears with the caspase gene family and the anti-apoptotic gene *BCL2*. Also, several critical nodes that entangle apoptosis and genome stability are cancer genes (e.g. *ATM*, *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6* and *TP53*), although their orthologs have arisen in different points of evolution. Our results demonstrate how genome stability and apoptosis were co-opted during evolution recruiting genes that merge both systems. We also provide several examples to exploit this evolutionary platform, where we have judiciously extended information on gene essentiality inferred from model organisms to human.

INTRODUCTION

The concept of apoptosis is associated with the maintenance of tissue homeostasis (1). The programmed cell death (PCD) in the perspective of multicellular organisms guarantees the substitution of old and/or dysfunctional cells, which are impaired by the accumulation of cellular damages due to environmental insults, as well as participates directly in tissue development (2). According to KEGG (3), a reference pathway database, there are up to 100 genes coordinately working in apoptosis. Removing one of these components affects several others and it may impair the whole pathway. In complex metazoan organisms, a defective apoptosis is associated with organogenesis disorders and also uncontrolled cell growth, which is typically found in neoplastic diseases (4). In the perspective of a cancer cell, suppressed apoptosis is a requirement in order to enhance cell fitness (5). In some extent, it is thought that apoptosis is related to genome instability in the sense that mutation prone clones, containing aberrant genetic content (i.e. high number of chromosome aberrations and DNA point-mutations), need a defective apoptosis to escape cell death (6–8).

Genome-maintenance mechanisms are intimately linked to apoptotic components, as indicates the high number of proteins that interact with the tumor-suppressor protein p53. In fact, this protein interacts with the four major DNA repair mechanisms: nucleotide excision repair (NER), base excision repair (BER), mismatch repair (MMR) and recombinational repair (RER)—homologous recombinational repair (HRR) and nonhomologous end-joining (NHEJ). Concerning NER and MMR, p53 can act in both transactivation-independent and -dependent manner (9). Furthermore, several DNA repair proteins can stimulate apoptosis in response to DNA lesions,

*To whom correspondence should be addressed. Tel: +55 51 3308 5577; Fax: +55 51 3308 5540; Email: mauro@ufrgs.br
Correspondence may also be addressed to Rita M.C. de Almeida. Tel: +55 51 3308 6521; Fax: +55 51 3308 7286; Email: rita@if.ufrgs.br

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

as for example the BER-associated protein poly(ADP-ribose) polymerase-1 (PARP1) (10) and the MMR proteins MSH2, MSH6 and MLH1 (11). Indeed, the overlapping among apoptosis and DNA repair genes renders difficult a precise definition of functional boundaries among all systems, which is a characteristic of complex biological networks (12).

On the other hand, apoptosis and genome-stability networks have different evolutionary roots. For instance, the core machinery of eukaryotic repair systems seems to be conserved among the three domains of life, although an expressive number of eukaryotic proteins have no counterpart in archaea or bacteria (13). Likewise, metazoan apoptosis contains several components that can be identified in ancient organisms such as prokaryotes and unicellular eukaryotes. However, many molecular sources in the eukaryotic apoptosis network might have been inherited from prokaryotes by horizontal gene transfer (HGT) in different events, being exapted to new functions to form apoptosis network (14).

Notwithstanding the components of these two networks having been extensively identified in eukaryote phylogeny (15,16), few data are available about the evolutionary scenario that functionally linked apoptosis to genome-stability gene network (5,17,18). One approach to assess the role of each component in a given interacting network is through comparative genomics. Using well-studied models, as yeast and mouse, comparative genomics provides powerful tools to draw evolutionary inferences for poorly studied organisms (16).

In a previous paper we characterized the entanglement among apoptosis and genome-stability pathways in a human protein–protein-association network (19). Here, we extend this characterization to build a platform to transfer functional information from several organisms to human. The idea is based on the consensus that each component of a gene/protein interaction network in the present living organisms has its origin at some point of the evolution. Thus the scenario that gives rise to the present network can be tracked-down by searching the root of each component in a given species tree.

Our goal here is to create an orthology map across a species tree for the human apoptosis and genome-stability gene/protein-association network in order to transfer to humans the information described for other eukaryotes. We searched for orthologs [i.e. homologous genes derived from a single ancestral gene in the last common ancestor (LCA) of compared species (20)] among 35 fully sequenced eukaryotic genomes. Likely orthology was inferred from orthologous groups using STRING database (21,22), and for each set of orthologs we found the most parsimonious scenario on the eukaryote phylogeny (23). To verify this orthology data, we reconstructed the entire analysis using Inparanoid database as a different data source, and essentially obtained the same results (see Supplementary materials). As further network characterizations, we estimated gene plasticity by measuring gene abundance and distribution of each orthologous groups among the extant species, and considered essentiality data available for yeast and mouse orthologs. Both plasticity and essentiality information were transferred to the human gene network. As a result we obtained a gene network where it is possible to discriminate ancient, less plastic and more essential regions from earlier, more plastic and less essential ones. Furthermore, the many cancer genes identified in this gene network are located in the earlier, more plastic and less essential region. We anticipate that our analyses can be applied to study the origins of a broad range of neoplastic diseases.

MATERIALS AND METHODS

Human gene/protein-association network

The protein–protein interaction network associating 180 human genes of apoptosis and genome-stability pathways has been extensively described in Ref. (19). Briefly, the network is generated using the database STRING (24) with input options 'databases','experiments' and 0.700 confidence level. STRING integrates different curated, public databases containing information on direct and indirect functional protein–protein associations. Each protein is identified according to both gene HUGO ID (25) and Ensembl Peptide ID (26) (Supplementary Table S1). The results from the search are saved in data files describing links between two genes and then handled in Medusa software (27).

Parsimony analysis: inferring evolutionary roots of human apoptosis and genome-stability genes

The parsimony analysis is divided into two major steps in order to construct parsimonious scenarios for individual sets of orthologous, given a species tree. We first built a consensus phylogeny for the eukaryotes listed in STRING database (22). The eukaryote phylogeny is based on a manual integration of a variety of phylogenies (28–33). We determined the presence of homologs among the organisms in the species tree for the 180 genes of apoptosis and genome-stability networks. Likely homology was inferred using the orthology information from the eukaryotic clusters of orthologous groups of proteins (KOGs) (21), which was retrieved through the orthology assignments in the STRING server; STRING has augmented the KOG orthology information by adding additional species (currently 35 eukaryotes) and creating more groups (NOGs, nonsupervised orthologous groups) as well as giving direct association among the three-domain phylogeny. In total, 142 eukaryotic orthologous groups were identified (Supplementary Table S1). To benchmark the analysis, we retrieved the orthologous groups for same set of genes using Inparanoid database, as discussed later.

The second major step is the reconstruction of the evolutionary scenario for each individual set of orthologous genes. This problem has been previously formulated as follows (23): given a species tree and a set of orthologs with a particular phyletic pattern, find the most parsimonious mapping for the set of orthologs on the tree. Precisely, concerning our problem, this question can be restated as: for each orthologous group associated with the human apoptosis and

genome-stability genes, find its earliest ortholog in the eukaryote phylogeny.

The incongruence of any evolutionary scenario is resolved according to the gain/penalty approach (23), where the most parsimonious scenario of presence/absence of all the genes at all ancestral nodes of the tree is obtained by using an inconsistency function defined as

$$S = \lambda + g\gamma, \quad \mathbf{1}$$

where λ is the number of gene losses, γ is the number of gene gains and g is the gain penalty. For each different scenario a function S is calculated and the most parsimonious scenario is chosen as the one that yields the minimum value of S. 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. This ratio is proposed by Mirkin and coworkers (23). Subsequently, other works validate the 2:1 ratio in prokaryotes (34,35) which thereafter has been used in similar analysis in eukaryotes and prokaryotes (36–38). Further details and the corresponding evolutionary scenario for all orthologous groups are presented in Supplementary Figures S14–S49 and also provided in spreadsheet format (Supplementary Table S3).

To verify the robustness of our orthology analysis we compared each gene evolutionary scenario with a corresponding one obtained using a different data source. In this case, we reconstructed the entire evolutionary analysis considering the Inparanoid database (39). 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. KOG and Inparanoid orthology analysis lead to roughly the same conclusions. We present and discuss these results in Supplementary Material Online (Supplementary Figures S3–S6, S50–S94 and Table S4).

Diversity analysis of orthologous groups

An orthologous group (OG) corresponds to a set of genes from different extant species that have a common gene ancestor. To obtain a quantitative expression of the orthologous distribution (i.e. distribution of the items of an orthologous group), we have measured the information content of two different databases (STRING and Inparanoid) using Shannon Information Theory (7,40–43) defined as follows. Consider n as the number of selected OGs, each one representing an orthologous groups. Each OG is labeled by α ($\alpha = 1, \ldots, n$) and has N_α items (orthologous genes), distributed among M possible organisms. Consequently, for a given OG we can define $s(i,\alpha)$ as being the number of items of a given organism i, ($i = 1, \ldots, M$), whose sum for a given α adds up to N_α The probability $p(\mathrm{i},\alpha)$ that, among the N_α items of the α-OG, a randomly chosen one belongs to the organism i is written as

$$p(i,\alpha) = \frac{s(i,\alpha)}{N_\alpha}, \quad \mathbf{2}$$

such that $\sum_i p(i,\alpha) = 1$. The normalized Shannon information function H_α is defined as

$$H_\alpha = -\frac{1}{\ln M}\sum_i p(i,\alpha)\ln p(i,\alpha), \quad \mathbf{3}$$

where we have divided by $\ln(M)$ in order to normalize the quantities, guaranteeing that $0 \le H_\alpha \le 1$. Observe that if there is one gene per organism, $N_\alpha = M$, $p(i,\alpha) = 1/M$, and $H_\alpha = 1$. In fact, H_α reflects the spread of the distribution $s(i,\alpha)$, i.e. it measures the diversity that exists in the αth OG. H_α near 0 indicates poor diversity, while a H_α close to 1 suggests high diversity. As a complementary quantity, we also estimate the abundance D_α in the αth OG by simply obtaining the ratio between the number of items (orthologous genes) and the number of organisms.

Transference of functional information from yeast and mouse to human gene/protein-association network

To predict developmental essentiality of a human gene, we used the mammalian phenotype information of the corresponding mouse orthologs. In this analysis, a gene is defined as 'essential' for organism development if a knock-out of a mouse ortholog confers embryonic or perinatal lethality (44). We obtained the mouse phenotype data from the curated knock-out collection available in Mouse Genome Database (MGD) (http://www.informatics.jax.org) (45). To predict cellular essentiality of a human gene, we used the phenotype information of the corresponding yeast orthologs. In this analysis, a human gene is defined as 'essential' at cellular level if a knock-out of its ortholog confers lethality to yeast. The yeast knock-out data were obtained from the *Saccharomyces* SGD project '*Saccharomyces* Genome Database' (http://www.yeastgenome.org/) (46). Human and yeast orthology is also verified using as data source the Inparanoid database (47) and is provided in Supplementary Table S1. In this analysis, six essential genes, out of 32, were not listed as orthologs when using Inparanoid (these genes are presented in Figure 5A with an asterisk besides their names).

Human gene mutation statistics

The data for the analysis of *CAN* genes is obtained from Cancer Gene Census (48). Both germline-mutated and somatic-mutated *CAN* genes are retrieved and then crossed with the list of 180 genes of our study. We identified 25 *CAN* genes placed in our network-based model of apoptosis and genome stability (Supplementary Table S1).

Genotype statistics of germline *CAN* genes located on ρ module is further analyzed in the XP mutation database (http://www.xpmutations.org). The representativeness of the sample was tested against a second database [Human gene Mutation Database—HGMD (49)] which is regarded as a reference mutation database for (published) gene lesions responsible for human inherited diseases. Table 1 shows as equivalent the samples obtained here from HGMD and XP database. However, the former contains limited gene information comparing to the latter (50).

Table 1. Allelic distribution of *CAN* genes placed in ρ module according to XP mutation database (Panel A). Sample representativeness compared to a second databases (Panel B)

Panel A	Number of Genotypes (%)[a]			Total genotypes	(Panel B) Entries[b]	
CAN gene	null/non-null	non-null/non-null	null/null		XP database	HGMD
ERCC2	20 (43.5)	26 (56.5)	0 (0.0)	46	76[c]	48
ERCC3	3 (60.0)	2 (40.0)	0 (0.0)	5	8	11
ERCC4	0 (0.0)	7 (100.0)	0 (0.0)	7	18[d]	17
ERCC5	0 (0.0)	5 (100.0)	0 (0.0)	5	10	12
XPA	6 (6.0)	94 (94.0)	0 (0.0)	100	128[e]	25
XPC	0 (0.0)	13 (100.0)	0 (0.0)	13	28[f]	42
DDB2	0 (0.0)	5 (100.0)	0 (0.0)	5	8[g]	8
Σ	29 (16.0)	153 (84.1)	0 (0.0)	182	276	163

[a]Data obtained from XP mutations database (http://www.xpmutations.org) is compiled according to the absence (null) or presence (non-null) of *CAN* gene alleles. Null/non-null genotypes are only heterozygous, while non-null/non-null genotypes include heterozygous and homozygous.
[b]The number of allelic records present in XP mutations database is compared to a second human inherited mutation database [Human gene Mutation Database — HGMD (49)] in order to attest the sample representativeness.
[c]One allele is duplicated in the database (the XP1BR entry).
[d]Three alleles have no mutation data (XP80TO, XP81TO and XP89TO entries).
[e]One allele had no zygosity information (XP10OS entry).
[f]Four alleles have no zygosity information (XP6BR, XP4BR, XP3BE and XP22BE entries). Polymorphisms are not considered in the analyses.
[g]One allele is duplicated (XP25PV entry).

Indeed, we could successfully retrieve the zygosity information only accessing the XP database.

RESULTS

Apoptosis and genome-stability gene set

Our analysis begins with a list of 180 genes participating in human apoptosis and genome-stability functions as previously defined (19) and provided as supplementary material online (Supplementary Table S1). To define this gene set we have characterized several genome-maintenance mechanisms as well as the interactions among their components. In Figure 1A we reproduce these interactions to illustrate the links between apoptosis and genome-stability gene networks, which are collectively referred to as the genome-maintenance gene network. Each node corresponds to a *gene-network node* (GNN), while the lines represent direct (physical) and/or indirect (functional) associations according to STRING database for human. They are derived from high-quality systematic protein–protein interaction mapping (22). Note the position of *TP53* gene in the network topology connecting apoptosis to 18 genome-stability components (Figure 1A, arrow, and Figure 1B). This functional overlap is further emphasized in Figure 1C for the complete network, which shows the number of links distributed for each gene set. Although apoptosis and genome stability have equivalent number of components in this network (i.e. 86:100), the connectivity of the latter is almost 2-fold, as indicated by the Venn diagram. Such difference arises mainly due to the large number of associations among NER, MMR and chromosome stability components, yielding a highly connected gene module (Figure 1A, ρ).

Construction of parsimonious evolutionary scenarios

In order to infer the ancestral states of human apoptosis and genome-stability genes we considered eukaryotic clusters of orthologous groups of proteins (KOGs) (21), using the orthology assignments in the STRING server (22). In total, apoptosis and genome-stability genes are distributed in 142 KOGs and for each one of these orthologous groups we found the most parsimonious mapping onto the eukaryote phylogeny. In Figure 2A we present the topology of the species tree used in this analysis (28–33), which is arranged in 17 subdivisions (monophyletic groups) based upon phylogenetic relationships. Every species-tree node (STN) is labeled according to the ascending subtree, and is referred to as the LCA of this subset.

To give a quantitative view of the evolutionary roots inferred for the 180 human genes studied here, we plotted the number of human apoptotic and genome-stability orthologs in each STN (Figure 2B). Accordingly, this distribution suggests a sequential enlargement of the network, with a progressive increase of apoptosis. In contrast, genome-stability orthologs are mainly rooted in STN-P (at the base of eukaryote species tree), suggesting that orthologs involved in apoptosis are more recent. Furthermore, in order to assess the robustness of our orthology analysis we reconstructed the entire evolutionary scenarios using Inparanoid database as a different data source, and essentially obtained the same results. 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 (39). We used this second approach to construct the evolutionary inconsistence score (R) that estimates the divergence between the two scenarios (i.e. ΔSTN). We present and discuss these results in Supplementary Material Online (Supplementary Figures S3–S6, S50–S94 and Supplementary Table S4). Briefly, for apoptosis genes, $R = 1.709$ STNs ± 0.224 (SE) and for genome-stability genes $R = 0.807$ STNs ± 0.202 (SE) (Figure 2C). It means that for each root inferred in our analyses, the estimated error for apoptosis is approximately two STNs up and down from the

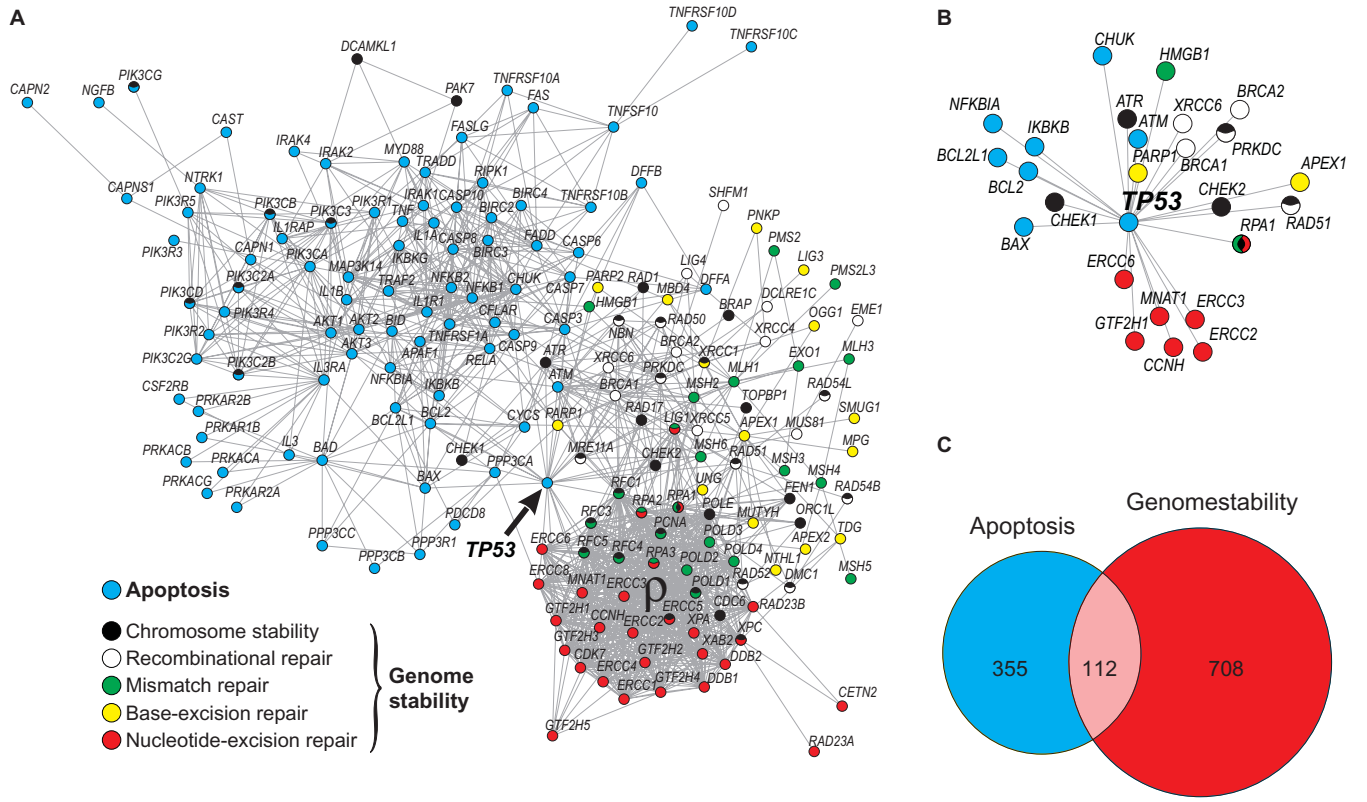


Figure 1. Human apoptosis and genome-stability gene network. (**A**) Graph of interactions among genes involved in apoptosis and DNA repair pathways, as previously characterized in Castro *et al.* (19). Different pathways are represented in different colors. Network nodes with more than one color represent genes participating in more than one pathway. Gene IDs of each pathway are provided in Supplementary Table S1. (**B**) Magnification of *TP53* gene position of in the network topology. It highlights the functional overlap of *TP53*, linking apoptosis to several genome-stability components. (**C**) Venn diagram showing the distribution of links between apoptosis and genome-stability pathways. The overlapped area corresponds to those links connecting both systems. The large number of associations among NER, MMR and chromosome-stability components is designed as ρ module.

rooting point in the species tree, while for genome stability the error is approximately one STN up and down.

In order to test a phylogeny where *Caenorhabditis elegans* is not at the root of the metazoa we included *Nematostella vectensis*, which thus changes the base of metazoa (Supplementary Figure S9). We chose this organism because (i) *Nematostella* is a cnidarian; (ii) the idea that the cnidarians are at the base of metazoa is less controversial than the nematodes; and (iii) switching a taxon like this goes some way to testing the effect of the phylogeny used. The result after this process is that the roots of the human genes remain almost the same (the complete analysis is available at Supplementary Table S5) and further discussed at supplements (section 1.4: the deep root of metazoans).

From species-tree nodes to gene-network nodes

To assess the details of the evolutionary scenario described earlier in the context of known and predicted gene functions, we used the network-based model presented in Figure 1A (19).

Starting from the complete network graph we generated three relevant orthology projections to characterize the functional differences between apoptosis and genome stability (Figure 3A–C). In these graphs we highlighted the nodes according to the roots inferred in the species tree (Figure 3D). Note that here each *gene-network node* (GNN) represents an ortholog of a gene in the human apoptosis and genome-stability gene network.

The orthology information regarding other STNs is provided in Supplementary Table S1. As quantitatively showed in Figures 2B–D, the more recent STNs concentrate apoptosis roots (round GNNs in Figure 3A and B). However, there is a qualitative difference: observe the pooled origins inferred for several components of apoptosis extrinsic (Figure 3A) and intrinsic (Figure 3B) pathways.

To analyze this result it is important to consider the biochemical signature of apoptosis, that is, the caspase activation, which is triggered by either intrinsic or extrinsic apoptosis pathways. The intrinsic pathway is associated with mitochondrial outer membrane permeabilization and cytochrome *c* (*CYCS*) release in response primarily to developmental cues or cellular damage. It triggers apoptosis through the Bcl-2 gene family and the initiator protease caspase-9. In contrast, the extrinsic pathway is characterized by the ligation of cell surface receptors via specific death ligands, as the *TNF* gene product, to generate catalytically active caspase-8 (51,52). The protein encoded by *TNF* gene is a multifunctional

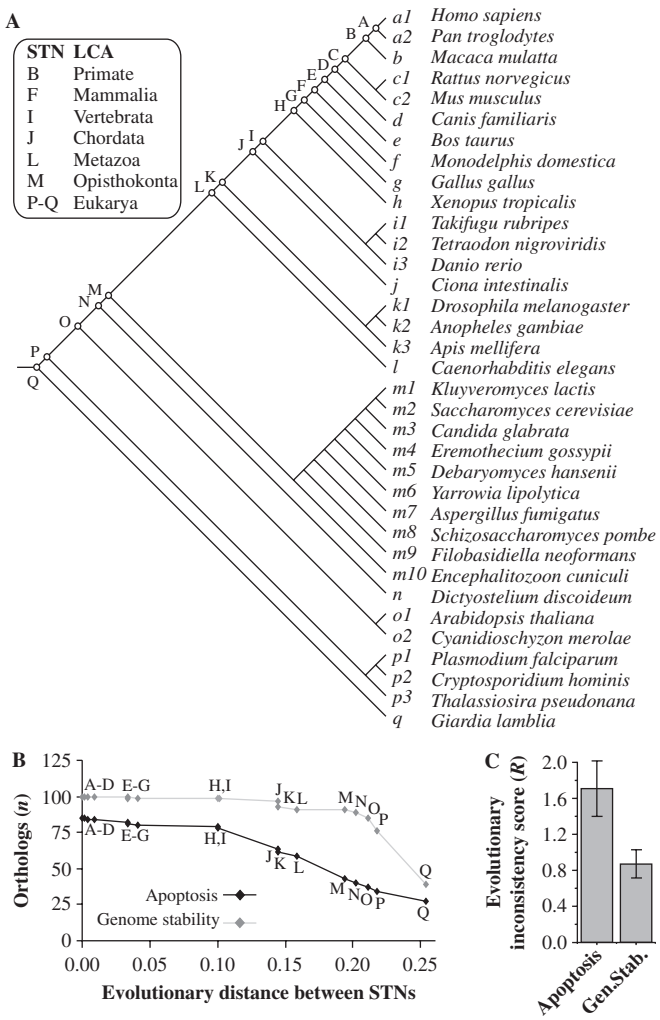


Figure 2. Inferring evolutionary roots of human apoptosis and genome-stability genes. (**A**) Eukaryote species tree topology used in the parsimony analysis. The phylogenic relationship among these 35 eukaryotes is based on a manual integration of a variety of phylogenies (28–33). STNs and the corresponding LCA are indicated. (**B**) Distribution of apoptotic and of genome-stability orthologs according to the roots inferred in the species tree and plotted as a function of the divergence between STNs (based on branch-length estimates). In Supplementary Material Online we exemplified the parsimony analysis. The evolutionary distances were computed using three protein families regarded as very conserved among distant taxa and described as able to reconstruct the three-domain phylogeny: 40S ribosomal proteins, translation initiation factor 5A proteins and Flap structure-specific endonuclease 1 proteins (73). All proteins used in the analysis are aligned in Supplementary Figures S10–S12. The distances are expressed as the fraction of sites that differ between the branches in a multiple alignment, which is an approximation of the branch-length that separates STNs. (**C**) Divergence between KOG and Inparanoid-derived scenarios. For apoptosis genes, $R = 1.709$ STNs ± 0.224 (SE) and for genome-stability genes $R = 0.807$ STNs ± 0.202 (SE). It means that for each root inferred in our analyses, the estimated error for apoptosis is approximately two STNs up and down from the rooting point in the species tree, while for genome stability the error is approximately one STN up and down.

proinflammatory cytokine that belongs to the tumor necrosis factor (TNF) superfamily, which also includes the ligands FAS (*FASLG*) and TRAIL (*TNFSF10*). These ligands bind to several members of TNF-receptor superfamily (e.g. *TNFRSF1A*, *TNFRSF10A*, *TNFRSF10B* and *FAS* receptors) and are involved in the regulation of a wide spectrum of biological processes, such as immune surveillance, innate immunity, haematopoiesis and tumor regression [for review, see (53)].

Accordingly, it is noticeable that the components of intrinsic pathway are rooted mainly in STN-L or earlier (e.g. *CYCS* is deeply rooted in eukaryote species tree—Figure 3C). In contrast, the subsequent enlargement of the network graph is provided mainly by orthologs of the extrinsic pathway, whose ligands and receptors are rooted in STN-I projection, or later (e.g. *IL1A*, *IL3RA*, *IL3* and *TNFRSF10D* genes are observed only in mammals, that is, STN-F and later, evinced by comparing STN-I projection versus complete human network; details of these orthologs are presented in the explicit parsimony analysis—Supplementary Figures S46, S48 and S49).

In STN-P projection (Figure 3C), however, only a small fraction of genes belongs to apoptosis. Instead, this graph is remarkable by the large presence of genome-stability components (triangular GNNs), as quantitatively addressed in Figure 2B.

Taking all results together, this evolutionary scenario of genome-maintenance mechanisms is marked by three

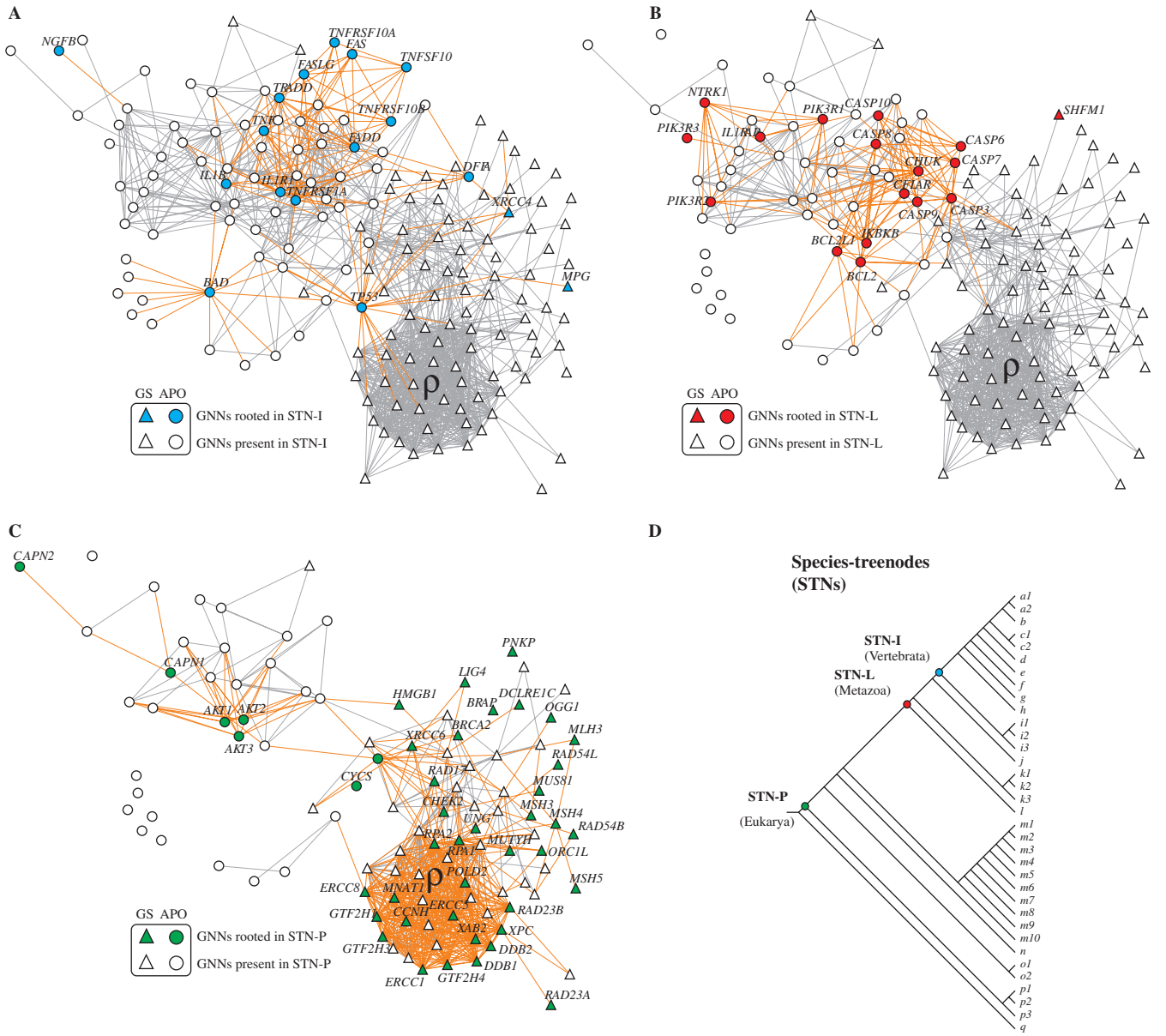


Figure 3. From STNs to gene-network nodes (GNNs). Orthology projection of genes rooted in STN-I (**A**), STN-L (**B**) and STN-P(**C**). Roots of an ortholog: color nodes; presence of an ortholog: white nodes. The location of these three STNs in the species tree is indicated (**D**). In Supplementary Material Online we provide further examples of this orthology projection approach (Supplementary Figure S2) and compared with Inparanoid evolutionary scenarios as a different orthology data source (Supplementary Figures S3–S6).

major functional increments: the first is the evolution of genome-stability gene network, whose components originate in the basal position of this species tree [STN-P, inconsistency between datasets $R = 0.63$ STNs ± 0.22 (SE)]; the second is the appearance of several apoptotic intrinsic components, rooted near metazoan divergence [STN-L, inconsistency between datasets $R = 1.23$ STNs ± 0.23 (SE)]; the third consists of the network enrichment with several apoptotic extrinsic members and happens near chordate-vertebrata root [STN-I, inconsistency between datasets $R = 0.35$ STNs ± 0.16 (SE)]. The network core of apoptosis and genome-stability systems are rooted in this tree before the divergence of metazoans, while GNNs placed in the periphery of the networks represent more recent evolutionary innovations. Therefore, the striking feature of these graphs is the increasing association between apoptosis and genome-stability functions with the emergence of an entangled gene network, which is fully consistent with the evolutionary strategy used in eukarya of adding complexity to existing core systems (54,55). (Inparanoid database essentially produces the same evolutionary scenario; please see Supplementary Figure S6.)

Also, additional evidence of the ancestral roots of genome stability 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 shows that 77.0% of the genome-stability orthologs have identifiable prokaryotic orthologous groups, against 39.5% for apoptotic orthologous genes.

Despite the several organisms that have been considered, the construction of the gene network is directed to human. Therefore, the interpretation of the evolutionary scenarios is ultimately linked with the characterization the human gene network. It means that we cannot infer that the gene network in the actual organism at the root of the eukaryotes was smaller. As we have stated in the introduction section, our goal is to create an orthology map across the species tree in order to transfer to human the information described for other eukaryotes. This is a one-way strategy, which is explored in the subsequent sections.

Plasticity analysis

Genetic plasticity may be understood as the ability of a functional gene network to tolerate changes in its components. There are different sources for such changes (gene duplication, gene loss, mutations and horizontal gene transfers), with different causes and effects. These changes in the genome may or may not be naturally selected, depending on the effect they have either on cell fitness or organism viability, in the case of complex organisms. The result of such an evolutionary dynamics is genetic variability among organisms of the same species or, ultimately, speciation. Gene networks are not equally plastic and hence do not equally respond to these variation pressures: depending on the gene, its function, influence on other genes, and their relevance, some changes are more likely to be tolerated or selected than others.

Focusing in networks in general, one may expect that gene networks that are more tolerant to variation will present a larger variability inside a species and among species. Focusing now on individual genes, organisms should be more tolerant to drastic changes (e.g. gene knock-out) when the change is performed on genes located at a more plastic network. These two characteristics, the gene variability among different genomes and the organism response to knock-out of single genes, allow two independent measures to estimate gene network plasticity. One possible plasticity measure is estimating the number and the distribution of orthologs among different organisms. A second, independent plasticity measure may be obtained by assessing cell lethality data. In what follows we present and discuss these two plasticity measures.

Diversity and abundance analysis. We evaluated the diversity and abundance of the orthologous groups to estimate the plasticity of each gene in our human apoptosis and genome-stability gene network (precise definition in the Materials and methods section and further exemplified in Supplementary Material Online).

The network graph presented in Figure 4A and B incorporates diversity and abundance statistics, allowing the discrimination in three distinct classes of genes based on the distribution of diversity as a function of abundance (Figure 4C). The first class (a) refers to genes placed in orthologous groups with low diversity and low abundance (Figure 4A and B, white GNNs; Figure 4C, white diamonds). It means that few organisms present these orthologs, and the associated orthologous groups have few components. This implies a very recent origin for these GNNs, since (i) all are present in humans, the end of our species tree; and (ii) they are not present in many extant species. For example, *TP53* and *FAS* have their origins at STN-I, as shown in Figures S38 and S40 in Supplementary Material Online. This class of genes must then be located at region of the network that is plastic enough to accept new genes. The second (b) refers to genes placed in orthologous groups with high diversity and low abundance (Figure 4A and B, black GNNs; Figure 4C, black diamonds), indicating a small number of genes per organism, but present in many different species. These genes are located in the most ancient region of the network. It implies poorly plastic genes, highly conserved among species. The last class (c) refer to those genes placed in orthologous groups with high diversity and high abundance (Figure 4A and B, red GNNs; Figure 4C, red diamonds), which clearly requires high plasticity. Note that both red and white GNNs (plastic GNNs) are segregated from the black GNNs (poorly plastic) in the network. This segregation should be expected since plasticity must be a characteristic of a set of interacting genes rather than a characteristic of an individual gene. Figure 4D supports these finding by showing the relative presence of the three classes of genes in the STNs: the more recent genes in the network emerge at the highly plastic regions of the network, while the more ancient ones are located at the poorly plastic regions.

Observe that this inhomogeneous distribution of white, red and black GNNs in the network graph reflects also in the function performed by the genes. While white and red GNNs are clearly populating apoptosis network, black GNNs are placed mainly in genome stability. This result suggests a high evolutionary conservation of genome-stability orthologs (i.e. class b, orthologs present in many organisms and with few variants), contrasting with apoptosis GNNs that concentrate the plasticity of the network (i.e. class c orthologs with many variants per organisms).

Essentiality in Saccharomyces cerevisiae. A second, independent plasticity measure is obtained by assessing cell lethality data. Here we considered the eukaryotic model *Saccharomyces cerevisiae* available in the *Saccharomyces* Genome Database (SGD) (46). We transferred this information to the STN representing the LCA of yeast and human (i.e. STN-M), which is then projected on the corresponding human network topology. The yeast results are showed in Figure 5A. Observe that essential genes are concentrated in a specific portion of the network (blue GNNs) corresponding to the lower plasticity area showed in Figure 4 (black GNNs there). Furthermore, likely orthology inferred in the LCA of yeast and human indicates that yeast have lost several genes in the course of its evolution, but mainly apoptotic genes (white GNNs in Figure 5A). Such loss, together with the presence of essential genes overlaid on genome-stability area

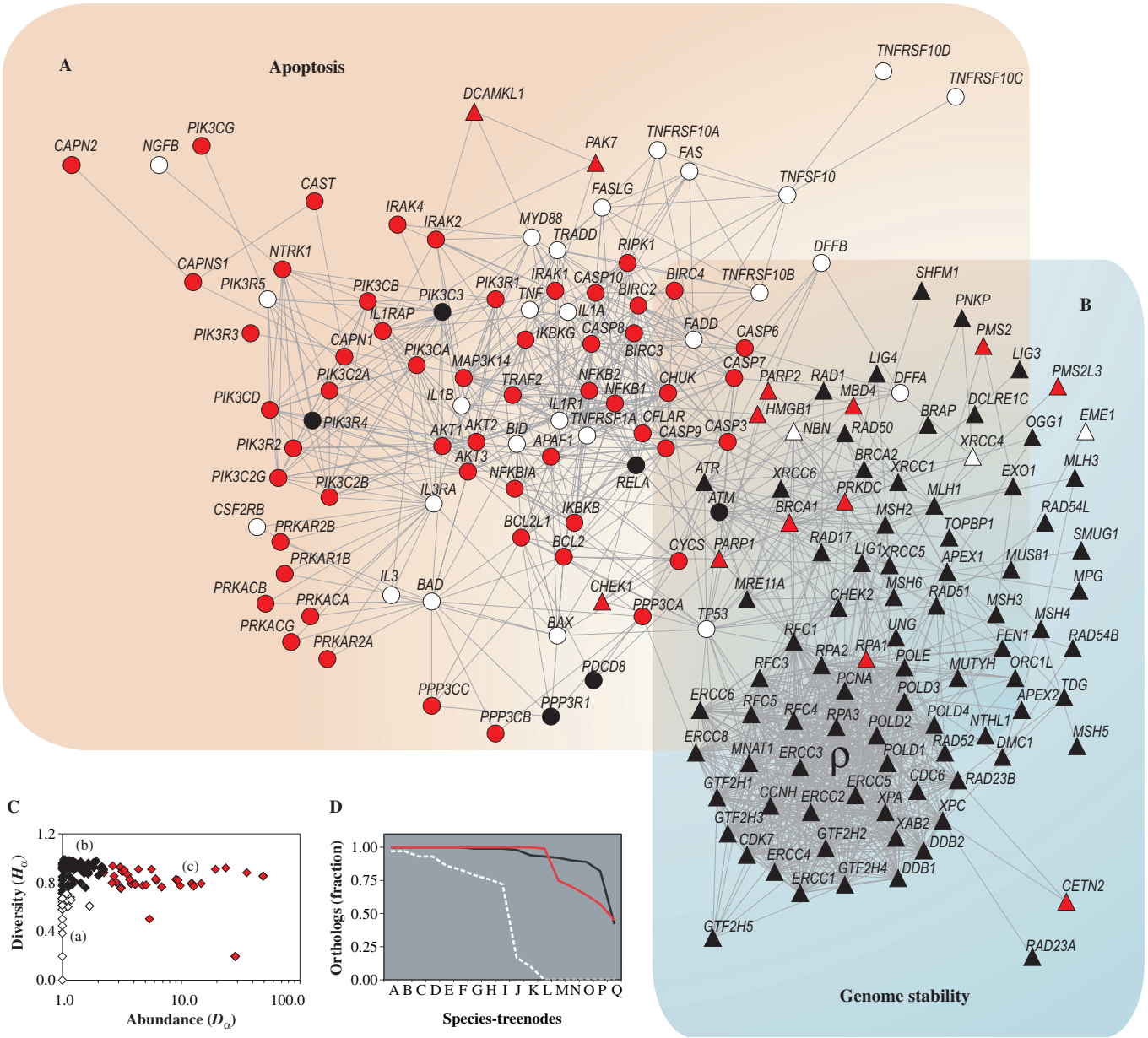


Figure 4. Plasticity analysis of orthologous groups. (**A** and **B**) Diversity H_α and abundance D_α of orthologous groups are overlaid on apoptosis and genome-stability gene network according to the categories defined in C. (**C**) Distribution of H_α as a function of D_α: (a) orthologous groups with low diversity and low abundance (white); (b) orthologous groups with high diversity and low abundance (black); (c) orthologous groups with high diversity and high abundance (red). (**D**) Fraction of orthologous groups present in the STNs: orthologous groups with low diversity and low abundance (white dashed line); orthologous groups with high diversity and low abundance (black solid line); and orthologous groups with high diversity and high abundance (red solid line). In Supplementary Material Online we provide examples of the diversity analysis.

(blue triangular GNNs), indicates that our evolutionary scenario is consistent with the plasticity measures shown in Figure 4: the lost genes are represented by plastic GNNs (red and white symbols in Figure 4).

Lethality in Mus musculus. In order to complement this lethality measure with a complex multicellular eukaryotic model, we assessed *Mus musculus* lethality data in Mouse Genome Database—MGD (45). 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. Evidence of mouse lethality is 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. Therefore, from the putative 178 *Mus musculus* orthologs identified in our analysis, we find 124 genes for which knock-out data are available (Supplementary Table S1). While the majority produced viable phenotypes, 39 knock-out alleles have been associated

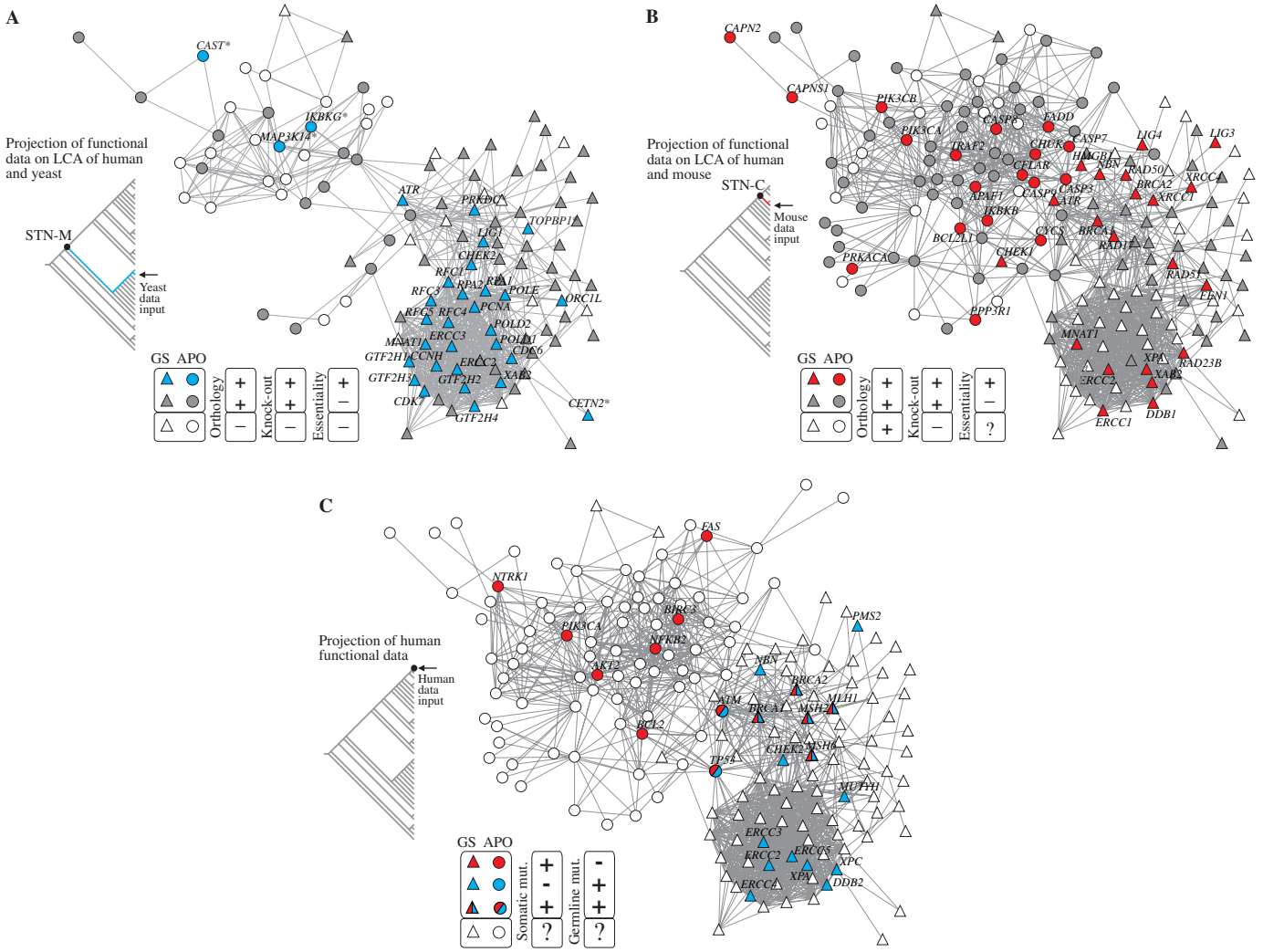


Figure 5. Integrating evolutionary and functional data. (**A**) Projection of yeast lethality data onto human apoptosis and genome-stability gene network: essential (blue GNNs) and nonessential yeast orthologs (grey GNNs) according to SGD database (46). The graph presents all orthologs inferred in the LCA of yeast and human (i.e. rooted or present in STN-M). White GNNs correspond to genes present in the branch but absent in yeast, as predicted in the parsimony analysis (see 'Materials and Methods' section). Asterisks identify six GNNs whose orthology are predicted by orthologous groups but not confirmed in the Inparanoid database (47). (**B**) Projection of mouse lethality data onto human apoptosis and genome-stability gene network: essential (red GNNs) and nonessential (grey GNNs) mouse orthologs according to MGD database (45). The graph presents only GNNs whose orthologs are inferred in the LCA of mouse and human (i.e. rooted or present in STN-C). GNNs that lack knock-out data in MGD database are indicated as white GNNs (mainly in ρ module). (**C**) Projection of genes causally implicated in human cancer—*CAN* genes—according to Cancer Gene Census (48). Colors indicate whether the gene is somatically mutated in cancer (red GNNs) or mutated in germline predisposing to cancer (blue GNNs) or both. White GNNs indicate genes not mentioned in the Cancer Gene Census.

with embryonic-perinatal lethality. The data are then transferred to the STN representing the LCA of mouse and human (i.e. STN-C) and then projected on the corresponding human network topology (Figure 5B). This data projection shows a homogeneous distribution of lethal alleles among nonlethal ones (red and grey GNNs, respectively), and a concentration on the genomic stability network of genes lacking knock-out data (white GNNs). Figure 5B highlights the essentiality of apoptosis and genome-stability gene network to the organism development. However, except for those genes without knock-out information (mainly placed in ρ module), mouse statistics indicate that the vast majority of knock-out alleles are nonessential at cellular level, given that even after gene disruption the cellular expansion is still viable. Such reading complements the results found for yeast, since what is nonessential to yeast is also nonessential to mouse at cellular level. A pictorial consequence of the complementarity of the results for yeast and mice is that the set of blue symbols in Figure 5A almost do not overlap with red symbols in Figure 5B.

Correlating plasticity and cancer statistics

The most systematical, available data about the functional impairment of human genome-maintenance mechanisms comes from cancer statistics. According to a global human disease network described by Goh *et al.* (44), from the 180 genes listed in our genome-maintenance

gene network, 51 are associated with some human disorder. From these, >50% are implicated in cancer. As an application for the plasticity estimates presented in the previous sections, we now consider cancer statistics data.

Genes causally implicated in cancer are collectively identified as cancer genes—*CAN* genes (48), and share a common feature: while they are potentially lethal to organism due to disruption of tissue architecture, mutations in these genes that lead to cancer are not lethal to the cell. These mutations are of two types: somatic or germline. While the first arise after organism development and in few cells, the second are inherited—present before conception—and thus continue afterwards in every cell. In fact, germline mutations in *CAN* genes cause cancer predisposition, not cancer *per se*, contrasting with somatic mutations that are to a large extent the primary cause of cancers (56).

Mutations that lead to cancer increase cell fitness (5,57), implying that the gene network may tolerate (and the cell may even benefit from) this genetic change (58). Consequently it is reasonable to expect that *CAN* genes are located on plastic gene networks.

We assessed the cancer statistics available in the Cancer Gene Census at the Cancer Genome Project—CGP (http://www.sanger.ac.uk/genetics/CGP). The graph of Figure 5C shows the projection of mutations causally implicated in human cancer retrieved from that census. Observe that *CAN* genes have a polarized distribution in the network topology. Those presenting exclusively somatic mutations are associated with apoptotic functions (red GNNs), and are at the plastic portion of the network, while those presenting exclusively germline ones are associated with genome stability (blue GNNs), at the poorly plastic region. Conversely, *CAN* genes that show both mutation types are at an in-between and overlap apoptosis and genome-stability networks.

The location of the germline mutations poses a challenge to our evolutionary scenario. How can we explain germline mutations in these human genes, given that they are located at a poorly plastic region? Also, care should be taken in order to consider these results together with yeast and mouse due to differences among statistical data. 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 (48). 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 (48). 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 is the most severe situation, equivalent to mouse and yeast knock-out data).

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, 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 turns reliable the obtaining of a representative sample (XP database is a repository of XP mutations identified in patients worldwide). We retrieved 182 mutated genotypes available in that database, which is then pooled according to the zygosity and the presence of null and nonnull alleles (Table 1, Panel A). Sample number is also compared to a second database in order to attest the representativeness of the database (Table 1, Panel B) (see Supplementary Material Online for further details). Given the data, in case null/null patients exist in some extent in human population, it would be a strong argument against the essentiality of genes located at the poorly plastic region of the network. As is pointed in Table 1, this is not the case. There is a total absence of null alleles in homozygous. Therefore, considering equivalent criteria among human, mouse and yeast to infer lethality, the data is consistent with lethality of germline *CAN* genes in the network projection, allowing the less-plastic area to be regarded as essential in human.

DISCUSSION

We presented an orthology map in order to locate the eukaryotic genes in the human apoptosis and genome-stability gene/protein-association network. According to our scenario, apoptosis and genome stability have different origins in the evolution, in spite of the complex interaction between both systems observed in human gene network (see Figure 6 for a summary). The genome-stability network seems to have emerged earlier in eukaryotic evolution.

Our results are consistent with several scenarios described by different authors. For instance, the position of genome stability in the base of eukaryotic species tree is highly consistence with the DNA repair functions described in prokaryotes [DNA repair in *Escherichia coli* is extensively recognized and has served as a paradigm for the investigation of other organisms: NER (59), BER (60), MMR (61) and RER (62)]. Also, the root of *BCL2* in the base of LCA of metazoans is consistent with the identified pro-survival functioning of Bcl-2 protein family members in *C. elegans* (63,64). Likewise, the position of caspases in the base of LCA of metazoans has been previously described (14), which is consistent with the origins of intrinsic pathway components that predate TNF-like cytokines (65). These TNF extrinsic pathway core components has been described across vertebrates (47) and corroborate our scenario, in line with the mammalian-like functioning

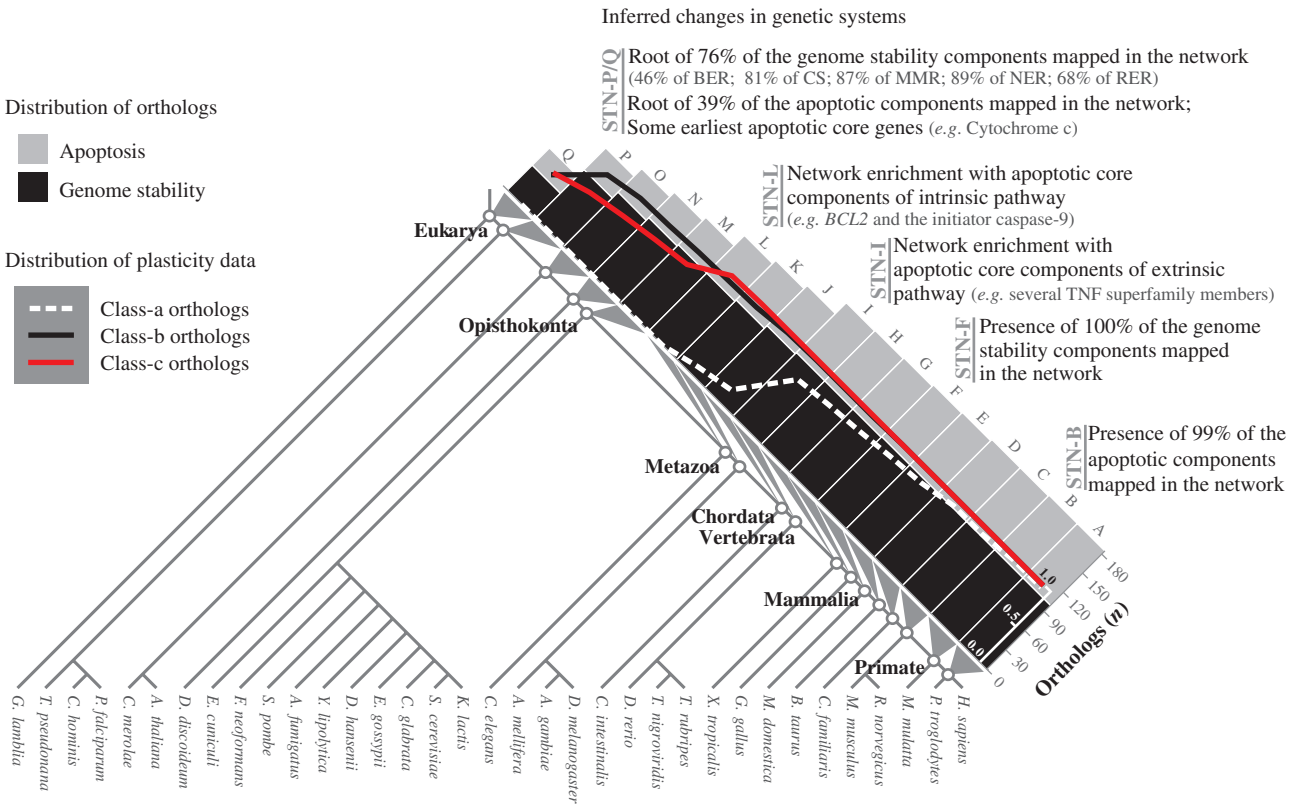


Figure 6. Summary of the inferred changes in genetic systems. The histograms show the distribution of 180 human orthologs according to the roots inferred in the eukaryote species tree (for details, see Figures 2 and 3). STNs and the corresponding LCA are indicated. Inset graph shows the presence fraction of orthologs of each STNs (for details, see Figure 4D). Diverse important events related to the roots of sets of genes are pointed along the STNs. Chromosome stability (CS).

of extrinsic apoptosis pathway described in *Danio rerio* and the absence of TNF and TNF receptor superfamily members in *C. elegans* (52).

However, the novelty here is that our results describe the genome-maintenance mechanisms as a whole, in a network-based model, to produce a unique evolutionary scenario. This point of view allows investigating the sequential events that led to the entanglement of apoptosis and genome-stability gene networks.

In the course of human genome-maintenance network evolution, three major functional increments are remarkable as is summarized in Figure 6. The first is associated to the base of the species tree and comprises genome-stability genes. The second evolves gradually, especially near the metazoan origin, with many gene components added to apoptosis intrinsic pathway, such as *BCL2* and the caspase gene family members. The third continues the apoptosis enrichment with the addition of several extrinsic components, such as TNF superfamily members.

Furthermore, as the macroevolutionary perspective of these conclusions must be considered together with the estimated evolutionary error (i.e. two species tree nodes up and down from the rooting point in the species tree), it is conceivable that some genes are actually not as recent as one might think. Nevertheless, our conclusions do point that in the course of human genome-maintenance gene network evolution there must have been a dramatic increase in the number of apoptotic components, contrasting with the early origin of genome-stability genes. We identified the expansion of apoptotic components in both KOG and Inparanoid-derived data.

This numerical expansion of apoptotic components could be related to the origin of other cell functions. Such assumption may be illustrated by the *TP53* appearance at the transition to later evolutionary scenarios: p53 protein regulates not only apoptosis, but it is also a key regulator of cellular senescence, defined as a permanent cell cycle arrest (66). Senescence is an alternative tumor suppressor mechanism, where damaged cells are prevented from dividing (67). If the senescence has functionally emerged with *TP53* gene, this second tumor-suppressor mechanism may have relaxed the selective pressure on apoptosis, increasing its tolerance against nonadaptive processes (e.g. genetic drift, mutation and recombination) and favoring its evolution. Our results are consistent with the emergence of both major mechanisms of tumor control during metazoan evolution, although in what regards senescence more genes should be taken into account to draw a safe conclusion.

Likewise, *TP53* can exemplify the evolution of genome-stability gene network. Acting as a transcription factor, p53 protein is able to modulate all DNA-repair processes (9,68). Such DNA-repair gene response to p53 protein is in line with evidences showing that even conserved gene functions are subject to substantial evolution at the regulatory level (69).

The plasticity analysis pointed the genes that during evolution suffered less duplication, such that they are poorly abundant and widely distributed among extant species. The results locate these more conserved genes mainly on the genome-stability network, which is also the more ancient portion of the network. In contrast, certain pairs of genes known to function together in human are placed in different distribution and abundance (e.g. *ATM* and *BRCA1*, *MUS81* and *EME1*, *PCNA* and *RPA1*, *RPA1* and *RPA2*—Figure 4). Analyzing together, it may indicate that the enlargement of the network can also occur through the addition of new nodes that eventually evolve to work together with ancient ones.

Lethality measures were performed in two complementary ways: one assessing knock-out data on yeast genes and the second regarding essentiality in mice. These two measures are complementary for the following reasons: yeast is a unicellular organism and lethality concerns only cell viability, while mouse is a multicellular animal, with a complex ontogeny. In this later case, a viable embryo implies survival after egg implantation and a relevant cell expansion. As a consequence, when an organism is labeled as viable, certainly the cell is viable and so is the organism. However, when the organism is not viable, the experimental procedure does not always discriminate whether the problem occurred at cell or at organism level. In summary, lethality data on unicellular organisms as yeast give sound information on what genes are essential for cell viability, while on multicellular organisms as mice the sound information is on what genes are not essential at cell level. Transferring cell essentiality information from mice and yeast to the human apoptosis and genome-stability gene network revealed that essential genes at cell level are mostly located at the more ancestral region of the network.

The integration of the information on ancestrality, plasticity and essentiality poses challenging questions. We found that the more ancient, less plastic and more essential genes are located on the genome stability, while the apoptosis network comprises the more recent, more plastic and less essential genes. Genome stability is required to guarantee the information transference from a parental genome to its offspring and thus provides one of the essential ingredients for natural selection to act: memory. It is not surprising that genome-stability network is rooted as early as possible in the species tree. It is also reasonable that such a crucial function is performed by highly conserved genes, where gene duplication is not favored due to the high possibility of disrupting a very essential pathway, yielding a poorly plastic network. Ancestrality, plasticity and essentiality have been pointed as correlated features in typical prokaryotes (70). On the other hand, in multicellular organisms with a more complex ontogeny, such as *Mus musculus*, the available literature reports not having found these correlations (71,72). Here we find cell gene essentiality to be correlated with ancestrality and plasticity in both unicellular and complex multicellular organisms. The point is that here we discriminate cell lethality from organism lethality: by isolating data from essential genes for cell survival from essential genes for organism viability, the correlation between cell essentiality, ancestrality and plasticity emerges and follows the same trends as in unicellular organisms.

A test for this putative evolutionary scenario for the human genome-maintenance network is given by the location of the human *CAN* genes. In more complex organisms natural selection acts at two different levels (organism fitness and cell fitness), what may stem conflicting selective pressures: while a fast proliferating cell clone is naturally selected in a unicellular organism, a fast proliferating cell clone in a complex organism may represent a tumor that may end up by killing the organism. In complex organisms, apoptosis and genome-stability networks work also as tissue-maintenance mechanisms, favoring natural selection acting at the organism level. As disruption of such a mechanism may favor natural selection acting at cell level, it stands to reason that many *CAN* genes are located at the plastic, less cell-essential region of the genome-maintenance network.

Specifically, concerning human functional data, at least two questions emerge from the evolutionary analysis of cancer statistics: (i) why the distribution of *CAN* genes is polarized between the two major segments described in the evolutionary scenario? and (ii) why *CAN* genes implicated in both types of cancers (somatic and germline) overlap apoptosis and genome-stability networks? While additional work will be needed to fully characterize the relevance of these results, it is clear for us that this evolutionary perspective may bring further insights in understanding cancer and its origins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank D. Jung for technical assistance. We acknowledge STRING, Inparanoid, MGD, SGD, CGP and XP databases for providing public access to their data.

FUNDING

Brazilian Agencies FAPERGS, CAPES and CNPq (grant 140947/2006-0, partially). Funding for open access charge: grant 40947/2006-0.

Conflict of interest statement. None declared.

REFERENCES

1. Danial,N.N. and Korsmeyer,S.J. (2004) Cell death: critical control points. *Cell*, **116**, 205–219.
2. Lettre,G. and Hengartner,M.O. (2006) Developmental apoptosis in C. elegans: a complex CEDnario. *Nat. Rev. Mol. Cell Biol.*, **7**, 97–108.
3. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
4. Hipfner,D.R. and Cohen,S.M. (2004) Connecting proliferation and apoptosis in development and disease. *Nat. Rev. Mol. Cell Biol.*, **5**, 805–815.

5. Crespi,B. and Summers,K. (2005) Evolutionary biology of cancer. *Trends Ecol. Evol.*, **20**, 545–552.
6. Yan,B., Wang,H., Peng,Y., Hu,Y., Wang,H., Zhang,X., Chen,Q., Bedford,J.S., Dewhirst,M.W. and Li,C.Y. (2006) A unique role of the DNA fragmentation factor in maintaining genomic stability. *Proc. Natl Acad. Sci. USA*, **103**, 1504–1509.
7. Castro,M.A.A., Onsten,T.G.H., Moreira,J.C.F. and de Almeida,R.M.C. (2006) Chromosome aberrations in solid tumors have a stochastic nature. *Mutat. Res.*, **600**, 150–164.
8. Zhivotovsky,B. and Kroemer,G. (2004) Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol.*, **5**, 752–762.
9. Sengupta,S. and Harris,C.C. (2005) p53: Traffic cop at the crossroads of DNA repair and recombination. *Nat. Rev. Mol. Cell Biol.*, **6**, 44–55.
10. Alano,C.C., Ying,W. and Swanson,R.A. (2004) Poly(ADP-ribose) polymerase-1-mediated cell death in astrocytes requires NAD+ depletion and mitochondrial permeability transition. *J. Biol. Chem.*, **279**, 18895–18902.
11. Duckett,D.R., Bronstein,S.M., Taya,Y. and Modrich,P. (1999) hMutSalpha- and hMutLalpha-dependent phosphorylation of p53 in response to DNA methylator damage. *Proc. Natl Acad. Sci. USA*, **96**, 12384–12388.
12. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
13. Aravind,L., Walker,D.R. and Koonin,E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.*, **27**, 1223–1242.
14. Koonin,E.V. and Aravind,L. (2002) Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ.*, **9**, 394–404.
15. Lin,Z., Kong,H., Nei,M. and Ma,H. (2006) Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc. Natl Acad. Sci. USA*, **103**, 10328–10333.
16. Aravind,L., Dixit,V.M. and Koonin,E.V. (2001) Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science*, **291**, 1279–1284.
17. Merlo,L.M.F., Pepper,J.W., Reid,B.J. and Maley,C.C. (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.
18. Greaves,M. (2007) Darwinian medicine: a case for cancer. *Nat. Rev. Cancer*, **7**, 213–221.
19. 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.
20. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
21. 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.
22. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
23. 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**, 2.
24. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
25. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.
26. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
27. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
28. 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.
29. Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.
30. Pennisi,E. (2003) Drafting a tree. *Science*, **300**, 1694.
31. Baldauf,S.L. (2003) The deep roots of eukaryotes. *Science*, **300**, 1703–1706.
32. 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.
33. Delsuc,F., Brinkmann,H. and Philippe,H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **6**, 361–375.
34. 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.
35. Kunin,V. and Ouzounis,C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res.*, **13**, 1589–1594.
36. Campillos,M., von Mering,C., Jensen,L.J. and Bork,P. (2006) Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res.*, **16**, 374–382.
37. Itoh,M., Nacher,J., Kuma,K.i, Goto,S. and Kanehisa,M. (2007) Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol.*, **8**, R121.
38. Pal,C., Papp,B. and Lercher,M.J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.*, **37**, 1372–1375.
39. 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.
40. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
41. Kendal,W.S. (1990) The use of information theory to analyze genomic changes in neoplasia. *Math. Biosci.*, **100**, 143–159.
42. Castro,M.A.A., Onsten,T.T.G., de Almeida,R.M.C. and Moreira,J.C.F. (2005) Profiling cytogenetic diversity with entropy-based karyotypic analysis. *J. Theor. Biol.*, **234**, 487–495.
43. Gatenby,R.A. and Frieden,B.R. (2004) Information dynamics in carcinogenesis and tumor growth. *Mutat. Res.*, **568**, 259–273.
44. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabasi,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
45. 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.
46. 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.
47. O'Brien,K.P., Remm,M. and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
48. 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.
49. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S.T., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human gene mutation database (HGMD (R)): 2003 update. *Hum. Mutat.*, **21**, 577–581.
50. Claustres,M., Horaitis,O., Vanevski,M. and Cotton,R.G.H. (2002) Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res.*, **12**, 680–688.
51. Beere,H.M. (2005) Death versus survival: functional interaction between the apoptotic and stress-inducible heat shock protein pathways. *J. Clin. Invest.*, **115**, 2633–2639.

52. Eimon,P.M., Kratz,E., Varfolomeev,E., Hymowitz,S.G., Stern,H., Zha,J. and Ashkenazi,A. (2006) Delineation of the cell-extrinsic apoptosis pathway in the zebrafish. *Cell Death Differ*., **13**, 1619–1630.
53. Aggarwal,B.B. (2003) Signalling pathways of the TNF superfamily: a double-edged sword. *Nat. Rev. Immunol*., **3**, 745–756.
54. 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.
55. Huettenbrenner,S., Maier,S., Leisser,C., Polgar,D., Strasser,S., Grusch,M. and Krupitza,G. (2003) The evolution of cell death programs as prerequisites of multicellularity. *Mutat. Res*., **543**, 235–249.
56. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med*., **10**, 789–799.
57. Breivik,J. and Gaudernack,G. (2004) Resolving the evolutionary paradox of genetic instability: a cost-benefit analysis of DNA repair in changing environments. *FEBS Lett*., **563**, 7–12.
58. Mombach,J.C., Castro,M.A., Moreira,J.C. and de Almeida,R.M. (2008) On the absence of mutations in nucleotide excision repair genes in sporadic solid tumors. *Genet. Mol. Res*., **7**, 152–160.
59. Setlow,R.B. and Carrier,W.L. (1964) Disappearance of thymine Dimers from DNA - error-correcting mechanism. *Proc. Natl Acad. Sci. USA*, **51**, 226–231.
60. Helling,R.B. (1968) Selection of a mutant of Escherichia coli which has high mutation rates. *J. Bacteriol*., **96**, 975–980.
61. Wildenberg,J. and Meselson,M. (1975) Mismatch repair in heteroduplex DNA. *Proc. Natl Acad. Sci. USA*, **72**, 2202–2206.
62. Willetts,N.S. and Clark,A.J. (1969) Characteristics of some multiply recombination-deficient strains of Escherichia coli. *J. Bacteriol*., **100**, 231–239.
63. Puthalakath,H. and Strasser,A. (2002) Keeping killers on a tight leash: transcriptional and posttranslational control of the pro-apoptotic activity of BH3-only proteins. *Cell Death Differ*., **9**, 505–512.
64. Youle,R.J. and Strasser,A. (2008) The BCL-2 protein family: opposing activities that mediate cell death. *Nat. Rev. Mol. Cell Biol*., **9**, 47–59.
65. Igaki,T., Kanda,H., Yamamoto-Goto,Y., Kanuka,H., Kuranaga,E., Aigaki,T. and Miura,M. (2002) Eiger, a TNF superfamily ligand that triggers the Drosophila JNK pathway. *EMBO J*., **21**, 3009–3018.
66. Rodier,F., Campisi,J. and Bhaumik,D. (2007) Two faces of p53: aging and tumor suppression. *Nucleic Acids Res*., **35**, 7475–7484.
67. Campisi,J. (2003) Cancer and ageing: rival demons? *Nat. Rev. Cancer*, **3**, 339–349.
68. Lavin,M.F. and Gueven,N. (2006) The complexity of p53 stabilization and activation. *Cell Death Differ*., **13**, 941–950.
69. Lynch,M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet*., **8**, 803–813.
70. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*., **12**, 962–968.
71. Liao,B.Y. and Zhang,J.Z. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet*., **23**, 378–381.
72. Liang,H. and Li,W.H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet*., **23**, 375–378.
73. 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.

Capítulo II

Evolutionary plasticity determination by orthologous groups distribution.

Artigo científico publicado no periódico *Biolgy Direct* (doi:10.1186/1745-6150-6-22).

BIOLOGY DIRECT

RESEARCH Open Access

Evolutionary plasticity determination by orthologous groups distribution

Rodrigo JS Dalmolin[1*], Mauro AA Castro[1], José L Rybarczyk Filho[1], Luis HT Souza[1], Rita MC de Almeida[2] and José CF Moreira[1]

Abstract

Background: Genetic plasticity may be understood as the ability of a functional gene network to tolerate alterations in its components or structure. Usually, the studies involving gene modifications in the course of the evolution are concerned to nucleotide sequence alterations in closely related species. However, the analysis of large scale data about the distribution of gene families in non-exclusively closely related species can provide insights on how plastic or how conserved a given gene family is. Here, we analyze the abundance and diversity of all Eukaryotic Clusters of Orthologous Groups (KOG) present in STRING database, resulting in a total of 4,850 KOGs. This dataset comprises 481,421 proteins distributed among 55 eukaryotes.

Results: We propose an index to evaluate the evolutionary plasticity and conservation of an orthologous group based on its abundance and diversity across eukaryotes. To further KOG plasticity analysis, we estimate the evolutionary distance average among all proteins which take part in the same orthologous group. As a result, we found a strong correlation between the evolutionary distance average and the proposed evolutionary plasticity index. Additionally, we found low evolutionary plasticity in *Saccharomyces cerevisiae* genes associated with inviability and *Mus musculus* genes associated with early lethality. At last, we plot the evolutionary plasticity value in different gene networks from yeast and humans. As a result, it was possible to discriminate among higher and lower plastic areas of the gene networks analyzed.

Conclusions: The distribution of gene families brings valuable information on evolutionary plasticity which might be related with genetic plasticity. Accordingly, it is possible to discriminate among conserved and plastic orthologous groups by evaluating their abundance and diversity across eukaryotes.

Reviewers: This article was reviewed by Prof Manyuan Long, Hiroyuki Toh, and Sebastien Halary.

Background

Biological systems are constantly changing at different hierarchical levels, such as genome sequences, gene/protein networks and organismal phenotypes. However, evolutionary constraints selectively act on all levels of organization allowing some changes and constraining others. Regarding specifically genomes, constraints do not act equally among all genetic sequences. Different classes of organisms (*e.g.* prokaryotes, unicellular eukaryotes, and multicellular eukaryotes) as well as different genomes structures (*e.g.* codifying sequences, introns, and "junk" sequences) can present huge differences in constraints. Even among codifying sequences, constraints act differently depending on the effect a possible mutation will generate on gene product. Synonymous mutations, for instance, are less constrained comparing to non-synonymous mutations. In addition, mutations in gene regions responsible for crucial sites, such as folding sites or enzymatic active sites, can be more constrained than disordered segments of proteins [1]. Considering genes as units, there are variable degrees of constraints leading to different evolutionary rates acting on different genes. Evolutionary rate of genes has been extensively studied, being related to several factors - not necessarily concurrent - such as gene expression level [2], gene essentiality [3], gene duplication [4], connectivity of the gene products [5], and gene age [6,7].

* Correspondence: rodrigo.dalmolin@ufrgs.br
[1]Department of Biochemistry, Institute of Basic Health Sciences, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil
Full list of author information is available at the end of the article

It is possible to describe the cellular metabolism by a graph or network, where gene or gene products are represented by nodes and their associations, by links. From the point of view of gene networks, genetic modifications might affect both links (interactions among gene products) and nodes (gene products). Modifications on genes structure, such as single mutation, deletions, or insertions can modify the interactions between the mutated gene product and its network partners (*e.g.* proteins participating in the same pathway), altering links of their network. Events as gene duplication and horizontal gene transfer modify the gene network by inserting nodes. In addition, network nodes can be deleted by gene loss events [8]. Similarly to genes, different gene networks might be subject to different constraints being more or less tolerant to changes and likewise presenting different levels of genetic plasticity - the ability of a functional gene or gene network to tolerate alterations in its components or structure [9].

Plasticity is an elusive property, in the sense it cannot be directly measured and it is always required a subjacent model to design a proper measure. Different artificial model networks have been proposed to define plasticity measures, bringing interesting conclusions on the possible functioning of biological networks [10,11]. In addition, *in silico* techniques have shown good power of prediction for metabolic networks in unicellular organisms [12,13]. In complex multicellular organisms, however, there is paucity of data. In effect, determining the plasticity of a given gene network is far from a straightforward task also due to the incomplete knowledge about the relationships among gene-products as well as about their behavior in different environmental conditions [14]. Regarding genes, a possible manner to experimentally investigate genetic plasticity is by using deletion analysis and different projects have developed and organized gene deletion information for different model organisms [15,16]. In this case, robustness against gene deletion may be interpreted as a tolerance against alterations on the network (node deletion), implying a correlation with plasticity. Deletion information is relatively well established for unicellular organisms such as yeast; for mammals, however, it involves more complicated and expensive techniques and the information is somewhat incomplete, even for model organisms.

A relevant problem one faces when defining a plasticity measure has to do with time scales. Here we consider time scales long enough to allow for speciation. For these time scales there is consensus that, for example, the nucleotide excision repair (NER) system is highly conserved: both the set of genes and the biochemical reactions they participate in are fairly similar in every extant eukaryote on Earth. Although this set of genes appeared very early in evolution, they have not been often deleted in descendent species and they have not suffered many duplications. Accordingly, each DNA repair genes has an ortholog in almost all species, without many paralogs [9]. Following this reasoning, we can infer that conserved, non-plastic genes belong to families spread over all eukaryotes with few paralogs. On the other hand, one could expect that ancient, plastic genes would have suffered deletions, and duplications in some species, but not in others, throughout evolutionary times. The consequence for their ortholog groups would be i) not having orthologs in many species, and ii) when a given species has a gene in those groups, they will also present many paralogous genes.

The crescent sea of data generated by genome sequencing projects has provided raw material to investigate the evolutionary relationships among genes from different species. The analysis of large scale data about the distribution of gene families (*i.e.* genes possessing the same common ancestor gene - an orthologous group [17]) across non-exclusively closely related species can provide insights about how plastic or how conserved a given orthologous group has been throughout its evolutionary history. In some extent, this *evolutionary plasticity* of an orthologous group might bring a perspective on the genetic plasticity of their orthologous genes. The idea is to estimate for each group of orthologs in eukaryotes the number of genes and how they are distributed among the species. From this information, properly processed, one can characterize their evolutionary history. For this measure to yield information, it must discriminate different orthologous groups. As shown in what follows, this is possible, since a considerable number of gene families has components spread in virtually all eukaryotes, whereas a great number of orthologous groups is restricted to some specific lineages [6]. Accordingly, the distribution analysis of a gene family in a species group brings valuable information about how conserved and how old that gene family is [7]. A common way to evaluate the breadth and the depth of a gene family distribution is based in looking for gene presence and absence in an evolutionary tree [18-20]. An alternative way to evaluate the distribution of an orthologous group consists in using the Shannon information theory [21] to determine the diversity (H_α) of its distribution in a species group [9]. This methodology is able to discriminate orthologous groups presenting patchy phylogenetic distributions - including lineage specific gene families - from broad distributed orthologous groups.

Molecular mechanisms such as gene duplication, exon shuffling, transposable elements, gene fusion and fission, and horizontal gene transfer have been related to development of new genes [22]. Among them, gene duplication has been discussed to be one of the most important

events in genome evolution by providing the prime source of genetic material in which evolutionary forces can act generating novelty [23,24]. Duplication events occur randomly and duplicated genes can address different fates: (*i*) they can be selectively preserved, mainly by bringing an adaptive advantage; (*ii*) they can be selectively eliminated by bringing an adaptive disadvantage; and (*iii*) they can remain unoccupied, drifting in evolutionary process, eventually being eliminated or, more rarely, evolving to develop another biological function [25]. It is noticeable some orthologous groups possess one-to-one relationships, while there are gene families composed by a great number of paralogs [26]. The reason why some duplicated genes are fixed while others are eliminated has been extensively discussed; however, the mechanisms driving the destiny of the new-born duplicated genes remain controversial [25,27-29]. The Neo-Functionalization (NEO-F) and the Escape from Adaptive Conflict (EAC) are among of the most important theories about the fixation of duplicated genes. NEO-F represents the first idea of evolution by gene duplication and suggests that once duplicated, one of the gene copies turns free to acquire a new function in the course of the accumulation of neutral mutations, while another copy preserves the original biological function. EAC suggests that a pleiotropic gene performing more than one function - where each function could not be independently improved - will be beneficed by a duplication event where each gene copy is then free to specialize in each different function former performed by a single gene. A third theory is represented by sub-functionalization, where degenerating mutations happens in both duplicated copies that subdivides gene function between the duplicated genes. Consequently, both altered copies are preserved by selection since any individual former gene is able to entirely perform their biological function (for review, see [29]). A useful method to identify the importance of duplication events in the evolutionary history of an orthologous group is given by the ratio between the number of components present in the orthologous group and the number of organisms containing items from this orthologous group.

In a previous paper, we analyzed the distribution and the duplicability of a set of 142 orthologous groups extracted from STRING database http://string.embl.de/ to investigate the evolutionary origin of human apoptosis and genome stability gene network [9]. Here, we extended the analysis to all Eukaryotic Clusters of Orthologous Groups (KOG) available in STRING. Our goal here is to evaluate the evolutionary plasticity and conservation of an orthologous group according to the distribution of their components (*i.e.* orthologous and paralogous proteins). For each KOG present in STRING database, we calculate the diversity and abundance of their components across 55 fully sequenced eukaryotic genomes and suggest an equation to determine the evolutionary plasticity taking into account both diversity and abundance. To further KOG plasticity analysis, we estimate the evolutionary distance average among all proteins which take part in the same orthologous group from a sample of the KOGs present in STRING database. As a result, we found a strong correlation between the evolutionary distance average and the evolutionary plasticity index proposed. Additionally, we evaluate the evolutionary plasticity of mouse and yeast genes associated with lethality when knocked-out. We found low evolutionary plasticity in *Saccharomyces cerevisiae* genes associated with inviability and *Mus musculus* genes associated with early lethality. At the end, we plot the evolutionary plasticity value in different gene networks from yeast and human to identify their more and less evolutionary plastic areas as well as their more and less evolutionary conserved areas.

Results

Genes distribution within Orthologous Groups

To assess the distribution of genes within each KOG we evaluated their diversity (H_α) and abundance (D_α) as described in *Methods* section. H_α provides the distribution of a given orthologous group across a species group. High diversity indicates an equalized distribution of KOG components (*i.e.* orthologous and paralogous proteins) among the species evaluated. On the contrary, low diversity suggests a non-homogenous distribution. For a KOG to present maximal diversity their components are present in all species, meaning that this KOG ancestral gene arrived early in evolution, in the last common ancestor of all considered organisms - in our case, in the origin of eukaryotes or before. Furthermore, besides this ancestral appearing early in evolution, for its descendants to be found in all assessed genomes, deletion episodes cannot have happened very often. D_α is defined as the average of number of proteins belonging to the same KOG, present in each organism. In general, high abundance denotes many duplication episodes in the evolutionary history of an orthologous group. Figure 1 shows the distribution according to H_α and D_α of all KOGs (4850 KOGs in total) present in STRING.

Note that there is a range of distribution, where H_α of the majority of the KOGs is around 0.8 to 1, while D_α is concentrated from 1 to 10. However, there are KOGs that show H_α values lower than 0.8 as well as KOGs that present D_α values higher than 10.

Evolutionary Plasticity Index

Low values of D_α combined with high values for H_α indicates low plastic orthologous group, since it is present in many species, with few components, indicating it

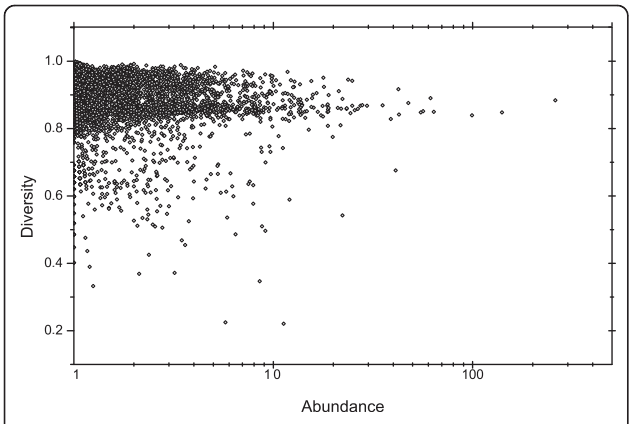


Figure 1 Diversity (H_α) and abundance (D_α) distribution. Each KOG present in STRING database was plot according to H_α(y axis) and D_α(x axis) values.

suffered few modifications (*i.e.* few duplication and deletion episodes) during eukaryotic evolution. Based on this, we have defined the evolutionary plastic index, *EPI*, to define how plastic a given orthologous group is, as follows:

$$EPI = 1 - \frac{H_\alpha}{\sqrt{D_\alpha}} . \quad (1)$$

Note that $0 \leq H_\alpha \leq 1$ and $D_\alpha \geq 1$. As a result, $0 \leq EPI \leq 1$. Figure 2A shows the distribution of all KOGs present in STRING organized in 100 groups according to *EPI*. Once identified the *EPI* of a given orthologous group, this information can be transferred to the proteins that compose this orthologous group (Figure 2B). The distribution of KOGs has its maximum displaced to low plasticity (Figure 2A); however, the distribution of proteins is roughly uniform (Figure 2B). This means that those KOGS with low plasticity present a lower number of proteins, strongly indicating a negative correlation between *EPI* and number of components (for further discussions, see Additional file 1, section 1.2).

Evolutionary Distance versus *EPI*

Genes can differ in their evolutionary rates. Genes under purifying selection evolve slower compared to genes under Darwinian selection [30]. In this sense, analyzing the amino acid differences among gene products from the same orthologous group might give us an alternative plasticity evaluation of a gene family. We compared the amino acid sequences, all against all, for a sample of KOGs present in STRING using Poisson correction method [31,32] as described in *Methods* section. This method analyzes the differences in amino acid sequences and provides an evolutionary distance between two proteins. We used the average of all distances among proteins of the same KOG to take the evolutionary distance average of each KOG evaluated. Note that we did not evaluate synonymous substitution since the analysis was performed utilizing amino acid sequences. Therefore, every observed difference corresponds to non-synonymous substitutions.

Figure 3A shows a strong correlation (Pearson correction 0.68621, two-tailed test $p < 0.0001$) between *EPI* and evolutionary distance of the evaluated KOGs. KOGs that possess high *EPI* present high evolutionary distance among their gene products as well as KOGs identified as having low *EPI* possess proteins more similar to each other. According to Figure 3A, the components of a KOG presenting low *EPI* are more similar among each other, comparing to components of a KOG presenting high *EPI*. No correlation was identified when plotting evolutionary distance versus D_α (Figure 2B), H_α (Figure 2C), number of species, and number of proteins (see Additional file 1, Supplementary Figure S6).

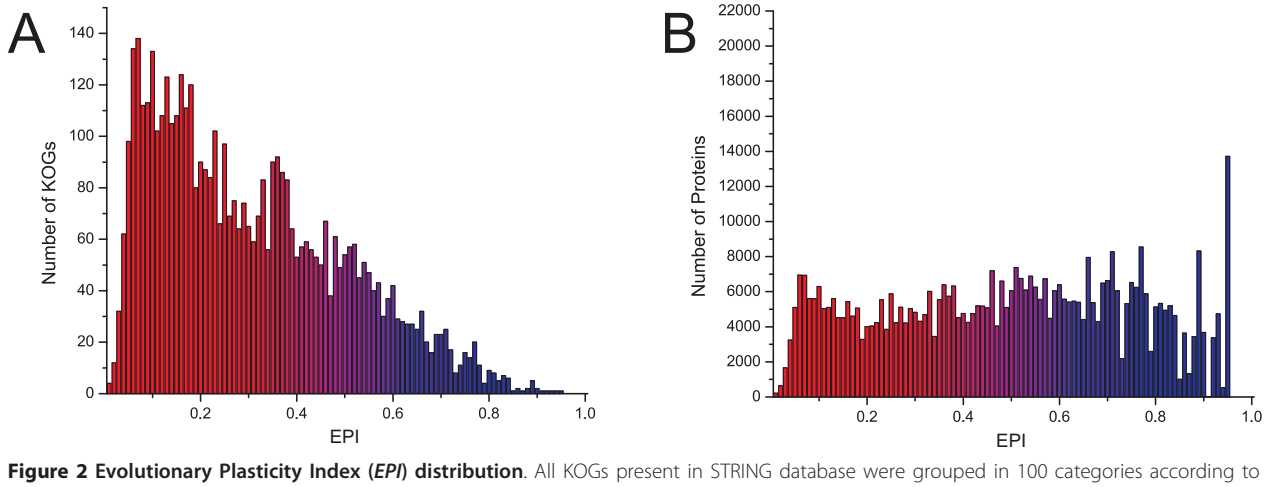


Figure 2 Evolutionary Plasticity Index (*EPI*) distribution. All KOGs present in STRING database were grouped in 100 categories according to *EPI* (A). All proteins present in KOG dataset were grouped in 100 categories according *EPI* (B).

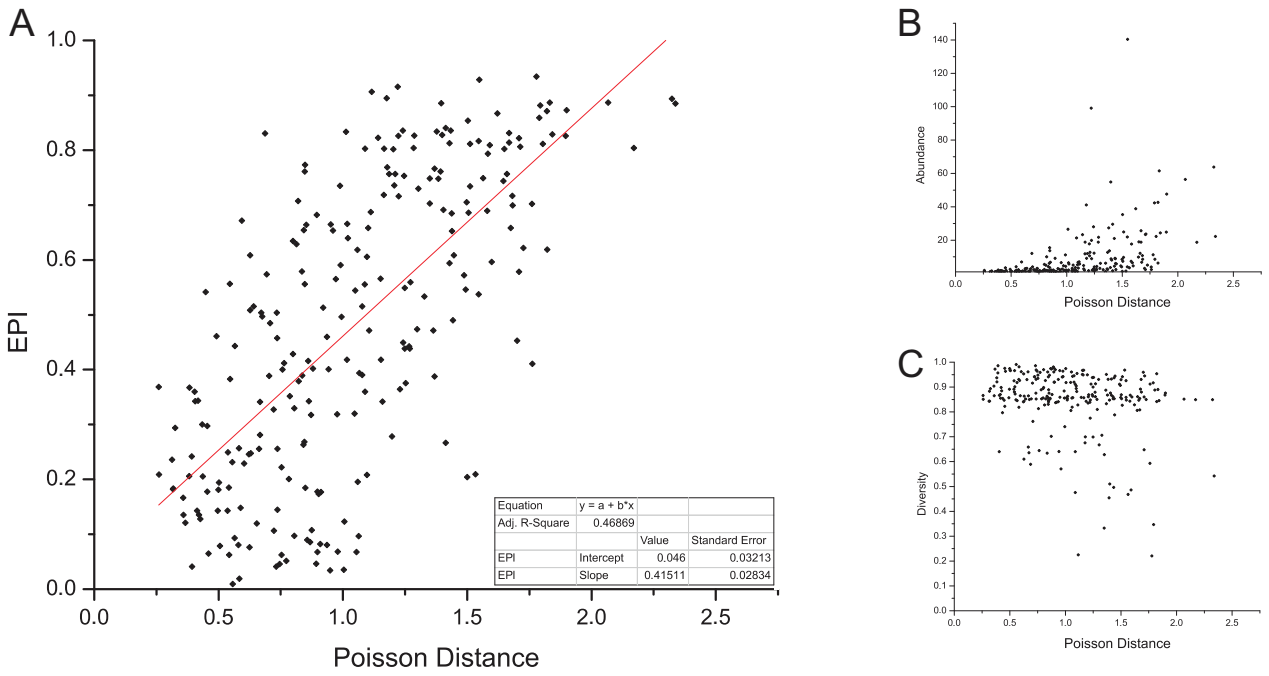


Figure 3 Evolutionary distance average versus *EPI*. 5% of the KOGs present in STRING database were sorted. The evolutionary distance among all proteins of each KOG evaluated was calculated and the evolutionary distance average (Poisson Distance) was obtained. Poisson Distance was plotted against *EPI* (A), abundance (B) and diversity (C) of each KOG evaluated. Red line indicates the linear regression fitting curve and the box shows the curve proprieties.

Functional Plasticity Analysis

To verify correlation of EPI with previous estimates of genetic plasticity, we assessed knock-out data from *Saccharomyces cerevisiae* and *Mus musculus*, and looked for genes related with lethality. We considered two criteria to identify genes involved with lethality: (*i*) *S. cerevisiae* genes which confer inviability when knocked-out and (*ii*) *M. musculus* target genes which cause early lethality (*i.e.* lethality before placentation). Additionally, we considered as viable *S. cerevisiae* genes annotated as "viable" in SGD as well as *M. musculus* genes annotated as "no abnormal phenotype detected" without any phenotype annotation associated with lethality in MGI (to further discussion, please see *Supplementary material*, section 1.3). Figure 4 shows the distribution of proteins from *S. cerevisiae* (Figure 4A) and *M. musculus* (Figure 4B) according to *EPI*. The grey landscape represents the *EPI* distribution of all proteins of *S. cerevisiae* (Figure 4A) and *M. musculus* (Figure 4B) present in KOG dataset. Yeast proteins present a distribution concentrated in low *EPI*, while mouse proteins present a more uniform *EPI* distribution (to further discussion, please see *Supplementary material*, section 1.4). The *EPI* distribution of proteins codified by genes involved with lethality when knocked-out have their maxima displaced to low *EPI* in both yeast and mouse (blue lines in Figures 4A and 4B, respectively). The opposite can be observed when considering proteins codified by genes associated to viable phenotype when knocked-out (red lines in Figures 4A and 4B). Figure 4C shows that mean *EPI* of inviable group is significantly lower comparing to mean *EPI* from all *S. cerevisiae* proteins present in KOG dataset. In the same way, the early lethality group has mean *EPI* significantly lower as compared to the totality of *M. musculus* proteins found in KOG dataset (Figure 4D). Additionally, mean *EPI* of viable groups are significantly higher when compared to respective total groups in both *S. cerevisiae* and *M. musculus* (Figure 4C and 4D, respectively).

Evolutionary Plasticity Index of biological networks

Cell functions are performed by functional modules [10,33] and gene network co-evolution has been proposed as an important evolutionary driving force agent [34]. In the same way, a network composed by proteins that take part in ancient and conserved KOGs can be regarded as conserved. To analyze the evolutionary plasticity of functional biological networks, we constructed the network of different pathways present in KEGG database http://www.genome.jp/kegg/ using protein interaction information from STRING (to further information, see *Methods* section). After network construction, we plotted the plasticity information of the network components (*i.e.* the *EPI* of the orthologous

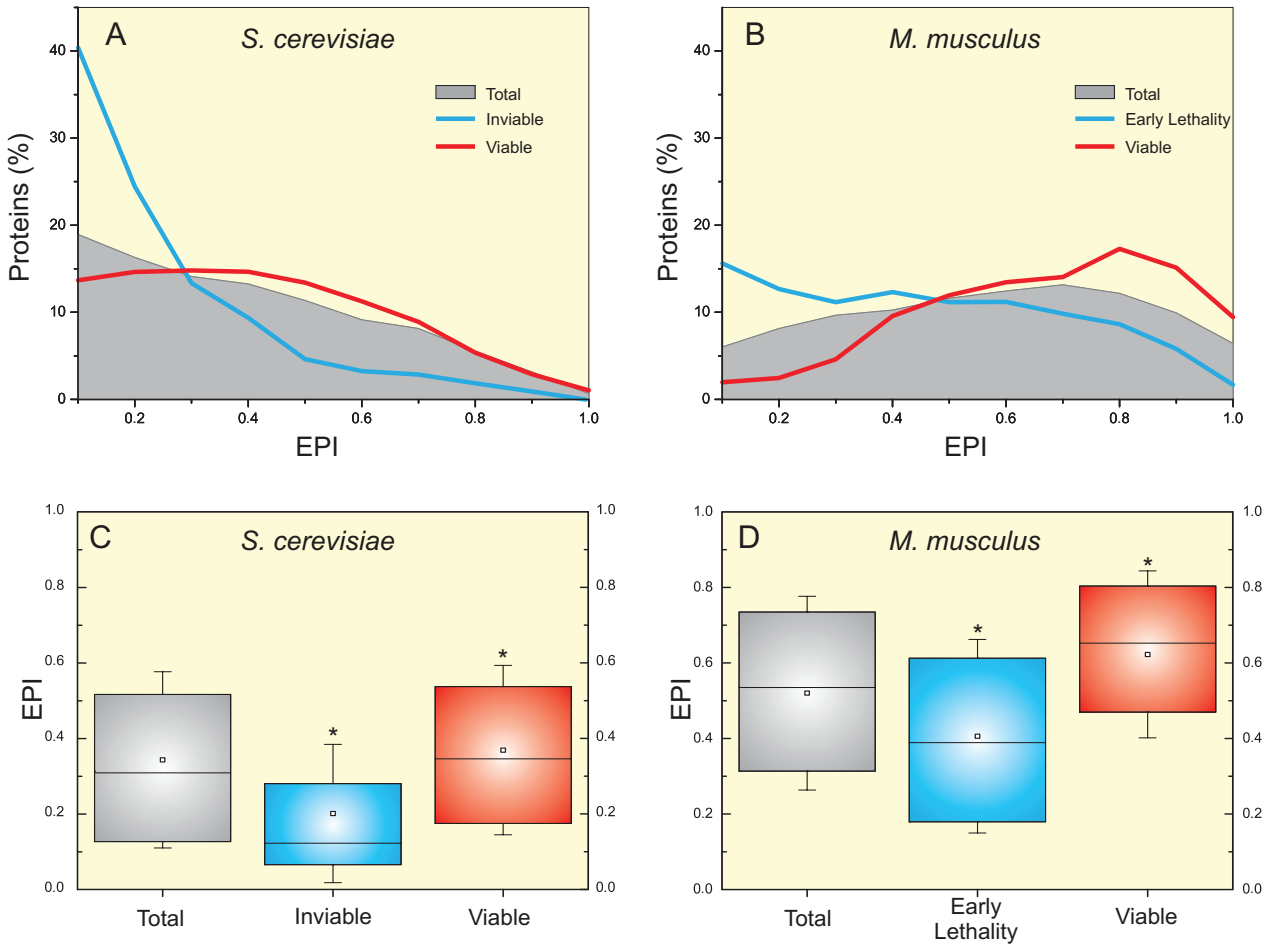


Figure 4 *EPI* distribution of target genes related with lethality when knocked-out. The percentage of *S. cerevisiae* (A) and *M. musculus* (B) genes presenting different *EPI* values is show. The grey landscape represents the *EPI* distribution of all genes from each species. Blue lines represent the *EPI* distribution of *S. cerevisiae* genes associated with inviable phenotype when knocked-out (A) and *M. musculus* target genes associated with early lethality (B). Red lines represent the *EPI* distribution of target genes associated with viable phenotypes (A and B). Boxes represent *EPI* distribution of the different gene categories (total, inviable, and viable) from *S. cerevisiae* (C) and (total, early lethality, and viable) from *M. musculus* (D). The edges of the boxes indicate the upper and lower quartiles. The line at the center of each box indicates the median, and the whiskers represent the standard deviation. * indicates different from total group ($p < 0.0001$).

group of each gene from the network) onto network topology. Figure 5 shows a graph representation of ribosome network from human (Figure 5A) and yeast (Figure 5C). Ribosome network is formed by a single highly connected module in both, human and yeast, and both networks present low evolutionary plasticity in their components (Figures 5B and 5D). Figures 5E and 5G show a graph representation for networks from several energetic pathways from human and yeast. Each network comprises components from glycolysis/gluconeogenesis metabolism, fatty acid metabolism, tricarboxylic acid (TCA) cycle, and oxidative phosphorylation. Differently from ribosome network, which is composed by one module, energetic metabolism network possesses several interconnected modules. As we can see in Figures 5F and 5H, the region comprising TCA cycle presents the lowest evolutionary plasticity in both human and yeast. Oxidative phosphorylation presents low, even though not the lowest, evolutionary plasticity and both, glycolysis/gluconeogenesis metabolism and fatty acid metabolism, present the highest evolutionary plasticity of human and yeast energetic metabolism network. Complete graph representation of the networks with gene symbols are available in Additional file 1 (Supplementary Figures S7, S8, and S9).

Discussion

Genetic plasticity estimative can be useful to different fields such as genetic diseases and evolution. For example, plasticity of a gene or a gene network can help finding components involved in pathology development as well as indicating possible therapeutical targets. Also, evolutionary

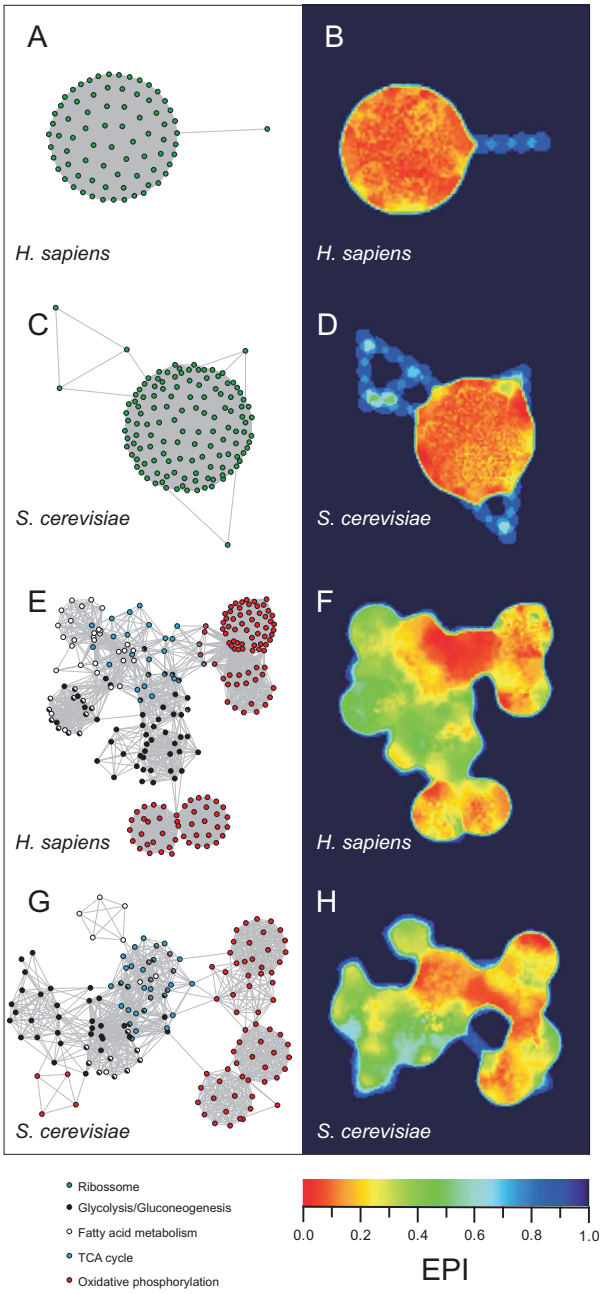


Figure 5 Gene networks *EPI*. *EPI* projection onto different graph representation of gene networks from *S. cerevisiae* and *H. sapiens* are shown, with the respective network topologies. The nodes represent genes and the links represent protein-protein interaction of gene products (A, C, E and F). The color landscape indicates the *EPI* (B, D, F, and H). The nodes were colored according to the pathways they belong (A, C, E and F). Nodes with more than one color belong to more than one pathway evaluated (E and G). To complete list of genes, please see Additional file 1 (*Supplementary Figures* S7, S8, and S9), Additional file 5 (*H. sapiens* genes) and Additional file 6 (*S. cerevisiae* genes).

novelty will probably appear on genome change-tolerant portions. The tolerance to modifications can be measured by directly modifying a gene structure or by estimating the gene variation in a population. Besides gene deletion experiments (a possible way of changing gene network structure), the presence of single-nucleotide polymorphism (SNP) (a way of estimating gene variation in a population) would be possible alternatives to evaluate genetic plasticity. However, a single nucleotide mutation may or not lead to a functional modification, depending on the site it occurs, leading to misevaluation of genetic plasticity. Copy number polymorphism (CNP) might work better in plasticity evaluation, mainly regarding entire deletions and duplication. In *Drosophila melanogaster*, for instance, around 8% of genes are at least partially duplicated and 2% are at least partially deleted, showing CNP as a common phenomenon and, consequently, an interesting target for genetic plasticity evaluation [35]. Genomic information has been largely used to predict biological function, from gene/protein function to entire gene/protein network architecture [14]. Co-inherence has been used to predict functional interaction between proteins [36] and computational techniques such as network alignment has been used to identify conserved pathways, manly in closely related organisms [37]. However, the evolutionary plasticity of orthologous groups has never been systematically analyzed.

Here, we have presented a large scale data analysis concerning the distribution of gene families across eukaryotes to identify conserved and plastic orthologous groups. It is noticeable the differences in orthologous groups distribution among eukaryotic genomes and those differences certainly hold biological information. The presence of a KOG component restricted to few eukaryotes indicate at least two possibilities: (*i*) the ancestral gene of this orthologous group arrived late in evolution and its orthologs are only observed in more recent taxa or (*ii*) the ancestral gene of this orthologous group arrived early, but its orthologs were lost in some of the taxa. Independently of the reason why a given orthologous group shows a patchy distribution among eukaryotes, it is clear that these orthologs are not required by all organisms. Conversely, a gene family widely found in eukaryotes plays an important role in virtually all organisms of this domain. Widely distributed genes have been described as being subject to stronger purifying selection as compared to young and less broadly distributed genes [6,7,18]. One hypothesis to explain these observations suggests that novel genes present an initial high evolutionary rate phase. At the end of this phase, there is a decrease in evolutionary rate

due to an increased functional constraint [6]. Recent works in *D. melanogaster* have shown an adaptive evolution of young genes and an increased purifying selection as genes become older, corroborating that hypothesis [38,39]. Therefore, genes belonging to essential ancient gene networks, which optimized their roles early in evolution, are expected to present high conserved components across a species tree as well as few drastic modifications in the course of their evolution. On the contrary, genes which arrived late in evolution - or even in ancient non-essential gene networks - might present a patchy distribution among eukaryotes.

Other important feature concerning orthologous groups is represented by gene duplication. Why some genes possess several paralogs whereas other genes maintain one-to-one orthology relationships? Despite gene duplication occurring randomly, some genes are prone to fix a duplication event while other genes avoid duplication. The fixation of a duplication event is commonly associated with function improvement in newborn duplicated copies. A very good example is given by Jones and Begun in their study involving three independent events of evolution of chimeric fusion genes in *D. melanogaster*. All three studied genes are derived *Adh* and all three genes experienced a rapid evolution on the beginning of their history, followed by a slower adaptive evolution. Additionally, the authors have observed an intriguing similarity in the pattern of evolution including temporal, spatial, and types of amino acid changes in these proteins [39]. Those data strongly suggest that the parent-protein characteristics might determine the path a possible copy will experience, including whether or not it will be fixed or eliminated. According to EAC theory, genes exercising more than one function (*i.e.* genes presenting functional plasticity) are prone to fix a possible duplication event [28,40].

EPI is based on drastic changes in the history of orthologous groups such as gene duplication and gene deletion. However, a gene may experience different degrees of changes. A gene highly tolerant to mutations will accumulate alterations in its nucleotide sequence on the course of its history. On the contrary, a gene lowly tolerant to mutations will present few nucleotide alterations in its evolutionary history. A complementary, independent measure of the plasticity of an orthologous group is then given by the similarity among the sequences of their proteins. Low evolutionary distances indicate that the proteins present very similar amino acid sequences. Consequently, they suffered few modifications as compared to those proteins presenting high evolutionary distance. According to our results, *EPI* is correlated to the evolutionary distance measure, suggesting that genes widely distributed among eukaryotes and possessing few paralogs are subject to purifying selection, reinforcing the idea that they are conserved, low plastic genes. A recent work involving gene families in primates has shown an interesting relationship among family size conservation, evolutionary rates and gene essentiality. According to the authors, genes within size conserved families present lower evolutionary rate and a higher proportion of essential genes compared to genes within non size conserved families from human, chimpanzee and rhesus [41]. Those results suggest that our observation concerning duplicability, diminished evolutionary rate, and increased essentiality can also be observed by analyzing gene families in closely related organisms.

The idea is not new that essential genes are subjected to stronger selective constraints and, consequently, evolve slower than nonessential genes [42]. In this sense, evolutionary plasticity could be the reflex of genetic plasticity. According to our results, genes associated with lethality are significantly more related to low plastic orthologous groups than genes associated with no abnormal phenotype in both *S. cerevisiae* and *M. musculus*. Therefore, the evolutionary history of a gene, *i.e.* the distribution of their orthologs among different organisms, might bring information about the relevance of their role. However, some less common exceptions may occur. It may happen that some new duplicated genes evolve to perform essential functions, as represented by essential genes with high *EPI*. Chen and collaborators have shown new genes that rapidly became essential in *D. melanogaster*, exercising crucial roles mainly in intermediary or late stages of development [38]. However, a wide distributed gene without duplication and deletion episodes probably exercise important biological role, suggesting that *EPI* may have more acuity to determine low plastic genes than to high plastic genes.

Our hypothesis that the evolutionary plasticity of an orthologous group can be an indicative of genetic plasticity of genes within that orthologous group has been applied to ribosome and energetic metabolism gene networks, showing interesting results. Ribosomes are known as ancient molecular fossils that have arrived before the LCA of all living organisms [43]. As it has been shown here, ribosome gene networks of both *S. cerevisiae* and *H. sapiens* present very low *EPI*. The entangled network topology indicates an intricate relationship among the partners of this very ancient low plastic gene network. On the other hand, central metabolism has been described as highly variable among different prokaryotes [44,45]. Here, we have found fatty acid metabolism and glycolysis/gluconeogenesis as the highest plastic portion in central metabolism. Despite glycolytic pathway might have arrived early in evolution, its components are not conserved across the species and glycolysis has been described as a high plastic and

versatile pathway [46]. Contrasting to glycolysis, TCA cycle represents the lowest *EPI* portion of the energetic metabolism network. Among the few works that have investigated the evolution of TCA cycle in eukaryotes, a recent paper has shown evolutionary similarity between mitochondria from *S. cerevisiae* and *Rickettsia prowazekii* in topological analyses based on network alignment and motif identification [47]. In the same work, the authors have described the mitochondria network as highly clustered around the TCA cycle. *R. prowazekii* is a mitochondria-related alpha-proteobacteria [48] and TCA cycle pathway seems to be closely related among eukaryotes and its ancestor prokaryote. Those assumptions agree with the results shown here, suggesting TCA cycle as low plastic and highly conserved among the eukaryotes. Despite the results shown here cannot be generalized to all biological networks, it opens a perspective on developing an extensive research concerning *EPI* and networks properties, such as node connectivity and clustering coefficient, as well as network centrality.

In the last decades, the advances in modern genomics have provided a powerful framework in the evolutionary research field. The availability of an enormous amount of completely sequenced genomes, including a great range of organisms, has provided new insights in evolutionary relationships involving genes, pathways, and species. *EPI* consists in a simple useful method that brings valuable complement in evolutionary studies and provides insights in other research fields such as pathology research and drug design. Clearly, the species set utilized in the orthologous group formation is essential to its diversity, abundance, and consequently to *EPI* determination. We avoid using the entire COG database due to its unequal distribution concerning the three domain of life (*i.e.* 532 *bacteria*, 43 *archaea* and 55 *eukarya*). *EPI* can be applied to any species group to identify the evolutionary plasticity of gene families in related species. However, the researcher must take care with the evolutionary relationship among the species used in *EPI* determination to avoid biased results. Many evolutionary questions, such as the exact factors determining gene and gene networks evolvability, are still unsolved. Despite our work does not clarify how and why the modifications of some gene networks are constrict on the course of evolution, *EPI* represents one step in evolutionary relationship understanding by identifying which gene families have been more or less stable on the course of evolution.

Conclusions

Our results suggest that the distribution of gene families brings valuable information on how plastic and how conserved a gene family is. It is possible to discriminate among conserved and plastic orthologous groups by evaluating their abundance and diversity. In addition, the evolutionary plasticity, measured according to orthologous group distribution as shown here, is coherent with other plasticity measures such as constriction in amino acid sequence modifications throughout evolution and essentiality in mouse and yeast. Finally, the evolutionary plasticity index measured according to abundance and diversity of gene families is consistent with the knowledge about the evolutionary conservation of ribosome gene network as well as the evolutionary plasticity of energetic metabolism gene network.

Methods

Data selection

Several databases offer tools in order to identify gene families. Each database utilizes its specific algorithm to find homology relationships according to specific purposes, such as to search orthologous genes/proteins along species or to search groups of genes/proteins which present the same last common ancestor (*i.e.* orthologous groups). However, the general strategy used by almost all database is to compare nucleotide sequences among different species [49-51] (to further discussion see Additional file 1, section 1.1). COG (Cluster of Orthologous Groups) database http://www.ncbi.nlm.nih.gov/COG presents a useful approach to identify orthologous groups. In COGs construction algorithm, all proteins encoded by the complete genomes analyzed are compared and for each protein, the best hit (BeT) in each different genome is detected. To name it as a cluster, it is necessary to form a triangle including BeT in at least three different organisms. Each COG represents a gene/protein family, including both orthologs and paralogs from different genomes, which have evolved from the same ancestral gene through a series of speciation and duplication events [52]. Besides COGs, which include eukaryotic and prokaryotic proteins, the database provides a tool involving only eukaryotic proteins. KOG (Eukaryotic Clusters of Orthologous Group) utilizes the same algorithm to find orthologous groups; however, it only works with eukaryotic genomes [53]. STRING database string-db.org has amplified the COG orthology information by creating more groups and adding extra species, totalizing 630 fully sequenced organisms with 55 eukaryotes among them [49]. Here, orthologous groups were accessed through STRING database version 8.2 stringdb.org [49], in download section. Only eukaryotic orthologous groups (KOG) were evaluated, resulting in a total of 4,850 KOGs. This dataset comprises 481,421 proteins distributed among 55 eukaryotes.

Distribution of orthologous groups

An orthologous group corresponds to a set of genes belonging to different species, which have a common

gene ancestor. To obtain a quantitative expression of the proteins distribution for each KOG (*i.e.* distribution of the items of a given KOG), we used Shannon Information Theory [9,21] defined as follows. Consider n as the number of selected KOGs, each one representing an orthologous group. Each KOG is labeled by α (α = 1,..., n) and has N_α items (orthologous and paralogous genes), distributed among M possible organisms. Consequently, for a given KOG we can define $s(i,\alpha)$ as the number of items of a given organism i, (i = 1,..., M_α), whose sum for a given α adds up to N_α. The probability p(i,α) that, among the N_α items of the α-KOG, a KOG randomly chosen, belongs to the organism i is written as

$$p(i,\alpha) = \frac{s(i,\alpha)}{N_\alpha} \quad (1a)$$

such that $\sum_i p(i,\alpha) = 1$. The normalized Shannon information function H_α is defined as

$$H_\alpha = -\frac{1}{\ln M} \sum_i p(i,\alpha) \ln p(i,\alpha) \quad (2)$$

where we have divided by ln(M) in order to normalize the quantities, guaranteeing that $0 \leq H_\alpha \leq 1$. Observe that if there is one gene per organism, $N_\alpha = M$, $p(i,\alpha) = 1/M$, and $H_\alpha = 1$. In fact, H_α reflects the spread of the distribution $s(i,\alpha)$, *i.e.*, it measures the diversity that exists in the α-th KOG. H_α near 0 indicates poor diversity, while a H_α close to 1 suggests high diversity. The abundance D_α of a given KOG was measured by obtaining the ratio between the number of items (orthologous and paralogous proteins) present in the KOG and the number of organisms containing items from this KOG. D_α vary from 1 to virtually infinite (despite the higher abundance found here was around 260) and represents the average of orthologous and paralogs per species for a given KOG. The diversity and abundance was conducted using the software GenPlast. GenPlast have been designed by our research group to perform the plasticity analysis presented in this paper. The software has been developed in the java platform, is under an open source license, and is freely available at http://lief.if.ufrgs.br/pub/biosoftwares/genplast.

Molecular evolutionary analysis

Molecular evolutionary analysis was conducted using MEGA version 4 [31]. 5% of the KOGs present in STRING (243 KOGs) was sorted according to the *EPI* and aligned amino acid sequences of all proteins comprising each KOG was obtained from STRING database string-db.org [49]. FASTA sequences were converted in MEGA format by the software. The number of amino acid substitutions per site between sequences was analyzed by the software set in "protein sequences". All results were based on the pairwise analysis sequences. Analyses were conducted using the Poisson correction method in MEGA4 [31,32]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). The Poisson Distance average of all proteins contained in each KOG evaluated was also obtained using MEGA4. To complete list of sorted KOGs, please see Additional file 2 (*Supplementary Table S4*).

Lethality Evaluation

Saccharomyces cerevisiae data was obtained from *Saccharomyces* Genome Database http://www.yeastgenome.org[16]. Genes associated to inviability when knocked-out was obtained using the SGD advanced search with step 1 (select chromosomal feature) set in "ORF" and step 2 (narrow results), box phenotype properties, set in "Inviable". Genes associated to viable phenotype when knocked-out was obtained following the same procedure, except by shift "inviable" by "viable" in phenotype properties box. The complete list of *S. cerevisiae* genes with phenotype annotations used here is available in Additional file 3 (*Supplementary Table* S5). *Mus musculus* data was obtained from Mouse Genome Informatics http://www.informatics.jax.org[54] in download area, file "Genotypes and Mammalian Phenotype Annotations (tab-delimited)". The following phenotype annotations were considered together to form the group "early lethality": embryonic lethality before implantation [MP:0006204], embryonic lethality at implantation [MP:0008527], embryonic lethality between implantation and placentation [MP:0009850], embryonic lethality before somite formation [MP:0006205], and embryonic lethality before turning of embryo [MP:0006206]. Genotypes with more than one target allele were discarded. Genes possessing the phenotype annotation "no abnormal phenotype detected [MP:0002169]" were considered to form the "viable" group. MGI BioMart version 0.7 http://biomart.informatics.jax.org was utilizing to find knock-out target genes. Genes combining MP:0002169 and any other phenotype annotation associated with lethality (MP:0005374, MP:0002081, MP:0002058, MP:0002080, MP:0008762, MP:0008527, MP:0006204, MP:0009850, MP:0006205, MP:0006206, MP:0006207, MP:0006208, MP:0005373, MP:0008569, and MP:0002082) were discarded. After that, all resultant genes codifying proteins preset in KOG dataset was utilized. The complete list of *M. musculus* genes with phenotype annotations used here is available in Additional file 4 (*Supplementary Table* S6).

Network Plasticity

The protein-protein interaction networks were generated using information from KEGG http://www.genome.jp/kegg/[51] and STRING string-db.org [49] databases

in two steps. First, the gene set of human and yeast pathways evaluated was obtained from KEGG. Only human genes identified in HUGO Gene Nomenclature Committee http://www.genenames.org/[55] and yeast genes identified in *Saccharomyces* Genome Database http://www.yeastgenome.org[56] were used. Second, protein interaction was obtained using STRING database with input options "databases", "experiments", and 0.700 confidence level. STRING integrates different curated public databases containing information on direct and indirect functional protein-protein associations. Network was constructed including only interacting genes/proteins and results from the search were saved and further handled in Medusa software [57]. Evolutionary plasticity of each network was determined in two steps. First, the *EPI* of each protein from the network was determined according to the *EPI* of the KOG to which the protein takes part. Second, the evolutionary plasticity data was plot onto the network using the software ViaComplex [58] to construct a landscape representation. The complete list of *H. sapiens* genes used to construct the networks is available in Additional file 5 (*Supplementary Tables S7*) and the complete list of *S. cerevisiae* genes used to construct the networks is available in Additional file 6 (*Supplementary table 8*).

Reviewers' comments

Reviewer 1

Professor Manyuan Long, Department of Ecology and Evolution The University of Chicago.

This reviewer provided no comments for publication.

Reviewer 2

Hiroyuki Toh,

The authors evaluated the evolutionary plasticity based on the diversity and the abundance of the orthlogous genes. The authors found that the plasticity is associated with the inviability of yeast and the early lethality of mouse. The approach is interesting. However, I found several problems in the manuscript. Following is the list for possible amendment.

Major: problems

(1) The authors defined "genetic plasticity" as the ability of a functional gene network to tolerate the alterations in its components or structures (p. 3 line 7 Background). In page 7 (Results, Evolutionary Plasticity Index), the authors defined "evolutionary plasticity" as formula (1), which is calculated with the diversity and the abundance of orthologous genes.

(1-1) Is the term "genetic plasticity" equivalent with the term "evolutionary plasticity"?

Authors' response: *Actually, these two concepts are different. Genetic plasticity is a gene property, while evolutionary plasticity is an orthologous group property. Genetic plasticity, as described on the manuscript, corresponds to the gene (or gene network) capacity to tolerate changes and the evolutionary plasticity, defined by Eq.1, is the record of changes a given gene family have experienced through its evolutionary history. We rewrite a substantial part of the introduction to clear both concepts. We also added additional discussion to elucidate the differences, as well as the relationships, between both.*

(1-2) If the two terms are used to indicate the same thing, it is not clear why the value calculated with the diversity and the abundance of the orthologous genes can indicate the ability of a gene network, since the formula (1) is given for a component of a gene network.

Authors' response: *As mentioned above, genetic plasticity and evolutionary plasticity are not the same thing. However, we propose a relationship between both. Starting to the point that genes do not work alone in an organism, the capacity of a gene to tolerate changes will certainly be influenced by their gene network. Additional discussions were added to the manuscript involving the relationship among the gene plasticity and the gene network plasticity.*

(2) about the term "paralog" used in the manuscript.

The authors used the eukaryotic clusters of orthologous group (KOG) in this study. To define the diversity and the abundance in p. 17 - 18 (Methods, Distribution of orthologous groups), the authors used not only orthologs but also paralogs. The description iseems to be confusing for the readers who are not so familiar with the genome science, since the term paralog" used in the manuscript is not the general one, I think that the authors wanted to indicate "co-ortholog" by "paralog". So, I think that a KOG does not include the distant paralogs. I recommend the authors to check the usage of the terms. Unless only close paralogs or co-orthologs are considered for the calculation of formula (1) in p. 18, the diversity loses the meanings. If the authors wanted to include the distant paralogs for the calculation, the consideration of the taxonomic bias may be required. Likewise, the definition of the abundance may be too naïve. Let's consider two cases with two species. In the first case, only one species has 99 paralogs, whereas the other has one orthologs. Da is calculated as (1+99)/2 = 50 in this case. In the other case, the first species has 50 paralogs and the other has remaining 50 copies. In this case, Da is calculated as (50 + 50)/2 = 50. That is the same values are obtained for the two cases. The first case may reflect a trend for the species specific gene amplification, whereas the second case may suggest the duplicability of the orthoologs. I think that the taxonomic bias should be taken into account for the calculation of Da.

Authors' response: *In fact, there are some controversies involving the terms ortholog and paralog. Other terms such as co-ortholog, inparalog, outparalog,*

pseudoortholog, pseudoparalog, etc, can be added to the debate. The strict description of each of those terms is not the point here. Our point is to discriminate among orthologous groups possessing one ortholog per species analyzed and orthologous groups possessing many orthologs (or co-orthologs) per species analyzed. Additionally, we analyze the distribution of ortholog among species to discriminate broadly distributed orthologous groups from poorly distributed orthologous group. As we do not include taxonomic relationships in the analysis, we analyze the set of species as a whole, independently of the distance among them. A fungi-specific orthologous group, for instance, will present low diversity. In the same way, a primate specific orthologous group also will present low diversity. In contrast, an orthologous group that has components equally present in all species evaluated will have high diversity. In what concerns KOG database, it intends to identify all eukaryotic genes which evolve from the same ancestral gene.

We agree with the reviewer in their comment relative to abundance. Abundance cannot be used without diversity to evolutionary plasticity inference, as shown in Figure 3B*. This is the reason why we use the abundance combined to diversity. Examining the suggested example:*

Case 1: one species has 99 paralogs, whereas the other has one ortholog. In this case, the abundance is 50 and the diversity is 0.080793136. Accordingly, EPI is 0.988574125.

Case 2: one species has 50 paralogs, whereas the other has 50 copies. The abundance is 50, exactly equal the case 1. The diversity, however, is 1. In this second case, EPI is 0.858578644. As shown, different orthologous groups presenting equal abundance but different diversity will have different EPI.

Minor problems

(1) The authors pointed out the importance of neofunctionalization after gene duplication. However, the authors did not mention sub-functionalization. I think that the dubfunctionalization is also related to the evolutionary plasticity. Why did the authors neglect the subfunctionalization.

Authors' response: *The theories discussed on the manuscript (i.e. neo-functionalization and EAC) are two important examples among many others about gene duplication theory. Since the reviewer judged important to mention sub-functionalization, a comment about that theory has been added on the manuscript.*

(2) The authors used "aminoacid" instead of amino acid in the manuscript. I think that "amino acid" is ordinarily used.

Authors' response: *It has been modified.*

(3) p. 19 -20 (Methods, Fitness Evaluation

The term "fitness" is used for different meanings from that in the evolutionary biology and the population genetics. I recommend the authors to use different term to express "fitness" in their manuscript.

Authors' response: *We have replaced "fitness" by "genes involved with lethality when knocked-out"*

(4) p.10 (Results, Functional Plasticity Analysis) naïve idea on evolution

It may be my misunderstanding, but some descriptions in p.10 seem to be naïve as an evolutionary statement.

(4-1),p. 10 line 1 "increase in complexity is a hallmark of evolution.

Evolutionary biologists do not consider so. Degeneration and neural change are also important to consider the evolution.

Authors' response: *We agree with the reviewer. In fact, there are examples of evolution by diminishing the complexity. The meaning intended with the sentence is related to life as a whole. Since first life forms have arrived, crescent levels of complexity can be observed in life organization. Despite simple organisms still represent the majority of the life forms, the complex relationships between different organisms and the environment is noticeable. To avoid misunderstanding we have changed "evolution" by "life" on the manuscript.*

(4-2) p. 10 lines 2 - 4

However, impairment in biological networks whose have arrived early in evolution (i.e. before multicellularity) might lead to early developmental lethality.

(5-1) whose ——□ which

Authors' response: *Alteration has been done.*

(5-2) There is no rationale or citation for this statement, but the authors seemed to follow the recaptulation theory by Heckel, which is still in debate. The authors should provide the rationale of this statement.

Authors' response: *We agree with the reviewer and removed the sentence from the main manuscript. We added an extra section in the Supplementary Material, discussing lethality in multicellular organism. We provide the rationale of that statement on this new section.*

Reviewer 3

Sebastien Halary,

Referee 3 - S. Halary

This study proposes an index called Evolutionary Plasticity Index (EPI) to assess the "genetic plasticity" of genes. This index is defined as a function of the abundance (number of genes) and distribution (diversity of organisms having these genes) within the homologous genes family a gene belongs to. EPI was calculated for 4850 KOGs and compared for 243 of them with their Poisson distance average of all proteins they contain. Then, EPI utility was illustrated by comparing the 'plasticity' of lethal against non-lethal genes of *S. cerevisiae*

and *M. musculus*, and the plasticity of genes involved in interactions/metabolic networks. EPI seems to be a simple tool to assess the diversity of a gene in eukaryotes, and then to be useful to characterize the paralogs richness of a homologous genes family. Nevertheless, this paper does not provide satisfactory arguments to justify the use of EPI rather than the other existing tools used up till now to estimate the diversity within a homologous family. This is mainly because the results are not discussed in sufficient depth. The authors propose to investigate relationships between EPI and lethality or topological position of the protein in a network, but did not compare their results with previous studies on the same subjects, whereas it could be useful to assess the power of their approach. To improve the manuscript, I would recommend that the authors provide concrete examples for which their index outperforms existing indices, or for which the tool is more straightforward. Also, the discussion can be improved by being more specific about optimal condition for this tool and/or by specifying novel applications.

From an editorial point of view, this paper is very long, mainly because of repetitions (without taking account of the 6 supplementary files). Many paragraphs are not placed in the suitable chapter. The quality of language could sometimes be improved upon as well. Overall, this results in a confusing article.

Authors' response: *We thank the reviewer for the extensive revision he had provided. We followed his suggestions as possible, improving substantially the paper. We also identify some misunderstanding and have worked on improve the clearness of the discussions.*

We agree with the reviewer and made efforts to make the paper as short as possible. We removed some peripheral discussions from the main manuscript to the supplementary files. We also have replaced many paragraphs in order to clear the reading. Language has been revised.

In my opinion, this article cannot be published before major editing and some revisions. I have some questions about the methods and the results, which I hope could be useful to improve the manuscript:

-How were the 5% of KOGs chosen for the comparison EPI/evolutionary distance? Why 5%?

Authors' response: *Data analyzed here involves a total of 481,421 proteins distributed among 4850 KOGs. It is a large, however finite, population. To better estimate the relationship among EPI and evolutionary distance, we take a large sample (i.e. n/N>0.05. In our case N = 4850. Accordingly, n would be > 242.5). A sample larger than 5% would be unnecessary and would substantially delay the paper.*

-There is a correlation between EPI and 'evolutionary distance', but it would be quite dangerous to resume the second by the first. These values provides more complementary than comparable information. You can find 2 KOGs with the same EPI, and very different means of distance (Figure 3A). Anyway, the authors discuss neither, nor do they comment on the relevance of their index. Which methods already exist to assess diversity of genes within a homologous family? Why is your index better than others or what kind of supplementary information can it provide?

Authors' response: *We completely agree with the reviewer. Evolutionary distance is complementary to EPI since both evaluate different classes of changes. While EPI identify entire gene alterations (i.e. duplication and deletion episodes), evolutionary plasticity evaluate the amino acid variation among the proteins. Since each measure evaluates different things, one cannot be explained exactly by a function of the other. Our results show that wide-distributed orthologous groups that have experienced few duplications and deletions episodes tend to have proteins more similar among each other (according to amino acid sequence), i.e. we found a coherent relationship between EPI and evolutionary distance, as shown by* Figure 3A. We have amplified the discussion about EPI and evolutionary distance relationship to clear it and to avoid misunderstanding.

Our analysis does not attempt to replace any existing method and the point here is the possibility to evaluate a great amount of data and extract information from it. The relationship among orthologs distribution and the orthologous group plasticity cannot be neglected and the present manuscript is the first work concerned in systematizing this relationship. We also have improved the discussion on other works concerning in evaluate gene networks plasticity to clear the usefulness of our research.

-The list of genomes in STRING DB (as I can read in the legend of Figure S5) is composed by 34 genomes from animals, 14 from fungi, 1 from plant and 6 from "protists" (belonging to 3 different kingdoms). First, for the figure S5A, if you choose to make the distinction between animals and fungi which are phylogenetically quite close. It could make sense to also make the distinction between the "protists" (mycetozoa, euglenozoa, alveolata and diplomonads) which are very distant from each other. Second, since there is just one plant in the dataset, you are not able to see the plant-specific KOGs and thus, you could underestimate the number of plant-specific paralogs and EPIs. Following the same reasoning, this study cannot be adapted to non-fungal unicellular organisms of the dataset. Actually, the dataset seems to be only suitable to assess animal and fungal protein diversity. What do you think about the possibility to adapt the set of genomes per study, to the organism of interest?

Authors' response: *The figure S5 attempts to show the EPI differences comparing complex multicellular and*

unicellular/simple multicellular organisms. Our intention is to discuss the relationship among the organism complexity (i.e. multicellular, unicellular) and EPI. The phylogenetic relationship among the groups is not the point here. We removed the figure S5A since we judge figure S5B as sufficient to discussion. Additionally, we add a new section on supplementary material to better discuss the results concerning EPI in different organisms. Regarding the second point, we agree with the reviewer. Species set is not appropriated to obtain conclusions on specific taxonomic groups such as plants. This is the reason way we do not infer any conclusion based on specific taxonomic groups. On the contrary, we just evaluated if an orthologous group is wide-distributed or narrowly-distributed among the 55 eukaryotes analyzed. We think is a good idea to use EPI to evaluate subsets of organisms and thank the referee for the suggestions. We add on the manuscript a discussion about this possibility.

-There are some "lethal proteins" with high EPI. Could you present one of these cases and discuss that?

Authors' response: *We have added an example of lethality in novel proteins of D. melanogaster. Additionally, we extend the discussion (supplemental material) regarding EPI and lethality.*

-You present lethal/non-lethal proteins study and network plasticity as two different cases of application, but it is probable, at least for some proteins, that their "lethality status" is related to their centrality and/or connectivity in the interactions network.

Authors' response: *We agree with and thank the reviewer for the suggestion. Indeed, many works have suggested an association among lethality and different networks properties, such as centrality and connectivity. Here, we found a connection between lethality and evolutionary plasticity, and a possible relationship among evolutionary plasticity, lethality, and networks properties may exist. One of ours perspectives is to perform a research involving evolutionary plasticity and networks properties.*

-You cite Li et al. 2006, which present the study of duplicability of genes in yeast. You could have cited also Chen et al 2010 (MBE) article which present a close investigation in humans. More precisely, I think you could have compared their results to yours to assess the efficiency and usefulness of your index before applying it to another question, even a simple one, to improve the quality of your discussion. You have focused your discussion exclusively on the importance of duplication in evolution, but you did not provide new evidence or hypotheses.

Authors' response: *We thank the reviewer to suggest the very good paper of Chen and collaborators. We have used their results in our discussion about duplicability and evolutionary rate.*

- It is a good idea to use your index to study diversity within metabolic/interaction networks. However, even if the 4 shown examples are interesting, they can not constitute any evidence about the importance of low EPI proteins within networks in general. A more convincing approach would have been to make an exhaustive study of protein's EPIs in function of their network's node properties. Centrality and connectivity measures should be useful to identify the proteins that you need to investigate in the aim to discuss about EAC theory, for instance.

Authors' response: *We completely agree with the reviewer. Our results must be evaluated as an example of EPI utilization. We have added a comment on discussion section to make it clear. As mentioned before, we plan to perform an extensive research involving evolutionary plasticity in a networks perspective.*

-The Figure 5 is very pretty, but it needs to be modified to improve the clarity of the results. First, (and at least) you must invert the both columns, since even in the text you began by describing the right one. Second, the resolution of the coloration in the left column is too low and it is often difficult to make the correlation between a node and its EPI. I propose to remove this column and to plot coloration directly on the network's nodes. For instance, Cytoscape allows to colorize a node and its outline in different colors. Furthermore, you don't provide a simple description of these networks in the legend and/or in Results: what are nodes, what are edges and what do the edges length mean?

Authors' response:

First: the columns have been inverted.

Second: we think may be a good strategy coloring the nodes to identify EPI values of the genes. However, it is not our objective here. The software ViaComplex, used to produce the figure, work by projecting a landscape onto a network to identify the area of influence of a given property, such as transcription level, lethality, or evolutionary plasticity. The software takes in consideration the nodes and the links between nodes to project the information (here, to project EPI information). To access EPI of a specific gene of the presented networks the reader can check supplementary tables S7 and S8.

Third: the figure brings a graph representation of different networks and the information regarding nodes and edges are presented on the figure legend. Additional figures with gene symbols are shown on supplementary material.

Please consider these detailed suggestions:

p.1: 2 semi-colons in the authors list.

Authors' response: *The commas have been substituted by semicolons in the author list.*

p. 2: in the Background section of the Abstract: "duplicability (abundance) and distribution (diversity)".

Is the abundance of genes in a COG (breadth of the COG) only a function of their duplicability? Can these 2 words be used as strictly synonyms.

Authors' response: *Every molecular mechanism involved with the development of new genes might be related to the abundance of an orthologous group. Horizontal gene transfer, for instance, can increase the abundance of an orthologous group by adding extra gene copies in a given genome, increasing the orthologous group abundance as a whole. Nevertheless, such episodes are far from greatly relevant in abundance constitution, mainly in eukaryotic organisms, whereas gene duplication is admittedly the most important mechanism. In addition, the exact molecular mechanism involved in abundance (i.e. gene duplication, reverse transcription, etc.) is not the point here. To avoid misunderstanding, however, we changed the referred sentence on the abstract.*

p. 3: "Genetic plasticity may be understood...". Exact repetition of the first sentence of the abstract.

Authors' response: *We have changed the sentence on the background section.*

p.3: «The analysis of a large scale data about the distribution of genes families (i.e orthologous group)». You did not survey families of orthologous genes *stricto sensu*, otherwise you should not have observed duplication events. Ortholog being a confusing term, especially when you use COGs from eggNOG database (which provides the db of STRING I think), it would be helpful to fix the definitions of homo/ortho/para-logous gene.

Authors' response: *We agree with the reviewer in their concernment about orthologs. It has been extensively discussed and there is no consensus about the nomenclature. The evolutionary relationships among genes involve several possible mechanisms that turn difficult to determine if a couple of genes in different species (or sometimes in the same species) are orthologs, coorthologs, paralogs, inparalogs, outparalogs, pseudoorthologs, or pseudoparalogs among each other. Despite such different relationships indeed exist, in practice, however, the identification and classification of homology relationships remains very difficult, mainly to entire genomes comparisons involving several species. The concept of orthologous group is exactly projected to characterize a group of genes with a same common ancestor, which is the meaning intended here. To make it clear, we have added this concept on the manuscript as well as the citation of a very explicative review wrote by Professor Koonin. Regarding the origin of the dataset, eggNOG and KOG represent distinct projects. KOG is based on a robust manual expert annotation whereas eggNOG is automatically and computationally constructed. For reference, please check Muller et al Nucl Acids Res 2010, 38: D190-D195.*

Then, the sentence p.4 « It is noticeable some orthologous groups possess one-to-one relationships, while there are gene families composed by a great number of paralogs» could be replaced by «Then, some homologous gene families are only composed by orthologs, while others possess a great number of paralogs too.»

Authors' response: *We think that to consider as orthologs all one-to-one relationship could be a mistake in some cases, according to discussed above. Let's examine the following example: There are two out-paralogs (gene A and gene A') in two related species (species x and species y). In this example, the gene A_x (i.e. the gene A from the species x) is ortholog of A_y (i.e. the gene A from the species y) and the gene A'_x is ortholog of A'_y. However, during the speciation process, the ortholog A has been deleted in a new species w (which possess only the gene A'_w) and the ortholog A' has been deleted in another new species k (which possesses only the gene A_k). Analyzing the species w and k, the genes A'_w and A_k are not orthologs among each other in spite of a one-to-one relationship involving the referred genes. Again, is very difficult to determine the exactly evolutionary relationship among genes. This is the reason way we prefer to use the orthologous group concept in our analysis.*

p.3: «from broad orthologous group» groups.

Authors' response: *The alteration has been done.*

p. 3 to p. 4: «In a previous paper, we analysed [...] to the genes which codifying such proteins. » These lines must be displaced in the last paragraph of the introduction. Furthermore, the syntax is not correct in « the genes which codifying such proteins ».

Authors' response: *The lines have been replaced and the last sentence has been removed.*

p. 6: «To assess the distribution of each KOGs», I would prefer «To assess the distribution of genes within each KOGs». In the same way, maybe you can change the title to make it more precise.

Authors' response: *The alterations have been done.*

p. 7: «As mentioned above, a KOG presenting $H\alpha$ and $D\alpha$ [...] few duplications episodes.» This sentence can be removed.

Authors' response: *The sentence has been removed.*

p. 7: «It is reasonable to think that a KOG with those [...] $H\alpha$ indicates a high plastic orthologous group.» More Discussion or Introduction than Results.

Authors' response: *The sentence has been modified.*

p. 7: «The distribution of KOGs is dislocated». Is "dislocated" the best term ?

Authors' response: *The term has been replaced.*

p. 8: «Accordingly, a randomly chosen protein has [...] characteristic of an index (to further discussion, see Additional file 1, section 1.2).» Discussion

Authors' response: *The sentence has been modified.*

p. 8: « Genes can differ according to evolutionary rates [...] plasticity evaluation of a gene family.» Discussion

Authors' response: *I agree with the reviewer that the discussion section would be a good place to the pointed sentence. However, we prefer provide a short introduction to situate the reader on the issue that will be presented. Additionally, the maintenance of the sentence will not disturb the objective of the section.*

p. 8: «We compared the aminoacid sequences [...] those proteins presenting high evolutionary distance.» Methods

Authors' response: *To the same reasons discussed above, we prefer to maintain a substantial part of the paragraph. The end of the paragraph, however, has been placed in the discussion section.*

p.8 to p. 9: «A gene highly tolerant to mutations [...] few nucleotide alterations in its evolutionary history.» Not Results.

Authors' response: *The sentence has been placed in the discussion section.*

p. 9: «That result suggests that genes widely distributed [...] they are conserved low plastic genes.» «Those results reinforce [...] than $D\alpha$ or $H\alpha$ individually.» Discussion. Furthermore, EPI better than $D\alpha$ or $H\alpha$, but is EPI better than 'evolutionnary distance' ???

Authors' response: *First, the sentences have been placed in the discussion section. Second, we did not make that statement regarding EPI better than evolutionary distance. Evolutionary distance is a measure of the divergence of amino acid sequence among proteins, i. e. it evaluates changes in protein's structures. EPI works with other kind of changes: gene duplications and gene deletions. So, they are complementary measures.*

p. 9: « Starting to the point that low [...] fitness impact when knocked-out...» to simplify.

Authors' response: *The paragraph has been rewritten.*

p.9: «*S. cerevisiae* information was obtained [...]Genome Informatics (MGI) [21].» Methods

Authors' response: *The sentence has been removed.*

p. 9: « Is not new the idea that [...] annotation associated with lethality in MGI.» Introduction/Discussion/Methods...not Results.

Authors' response: *A substantial part of the paragraph has been placed on discussion section.*

p.11: « To analyze the evolutionary plasticity [...] (*i.e.* the *EPI* of the orthologous group of each gene from the network) onto network topology. » Methods

Authors' response: *We believe that a brief introduction is important for a better presentation of the results of* Figure 5.

p. 12: «The evaluation of the history of a gene or a gene network is fundamental to understand its evolutionary behavior.» Do you mean that we need to know the evolution of a gene to understand its evolution? I think this sentence is not useful. Discussion must be revised. To improve the clarity of the discussion, you can follow the same structure than the Results chapter.

Authors' response: *The sentence has been removed. We substantially changed introduction and discussion sections.*

Methods must be simplified.

Authors' response: *Methods section has been simplified as possible.*

Additional material

Additional file 1: Supplementary Material. Document containing supplementary results and discussion, including 23 figures and 3 tables.

Additional file 2: Supplementary table S4. Table containing the orthologous groups sorted to evaluate the evolutionary distance among their proteins.

Additional file 3: Supplementary table S5. Table containing *Saccharomyces cerevisiae* genes possessing phenotype annotations involved with inviability or viability when knocked-out.

Additional file 4: Supplementary table S6. Table containing *Mus musculus* genes possessing phenotype annotations involved with "early lethality" or "no abnormal phenotype" when knocked-out.

Additional file 5: Supplementary table S7. Table containing *Homo sapiens* genes used to construct the networks to illustrate different biochemical pathways.

Additional file 6: Supplementary table S8. Table containing *Saccharomyces cerevisiae* genes used to construct the networks to illustrate different biochemical pathways.

Acknowledgements and Funding
We acknowledge STRING, KEGG, MGI, and SGD databases for providing public access to their data. This work was supported by PNPD SUS/CAPES.

Author details
[1]Department of Biochemistry, Institute of Basic Health Sciences, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil. [2]Department of Physics, Institute of Physics, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil.

Authors' contributions
RJSD designed the study, performed the analysis, discussed the results, and write the manuscript. MAAC designed the study and discussed the results. JLRF performed the analysis and discussed the results. LHTS performed the analysis. RMCA designed the study and critically reviewed the manuscript. JCFM designed the study and critically reviewed the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 19 January 2011 Accepted: 17 May 2011
Published: 17 May 2011

References

1. Koonin EV, Wolf YI: **Constraints and plasticity in genome and molecular-phenome evolution.** *Nat Rev Genet* 2010, **11**:487-498.
2. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**:927-931.
3. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:5483-5488.

4. Yang J, Gu Z, Li WH: **Rate of protein evolution versus fitness effect of gene deletion.** *Mol Biol Evol* 2003, **20**:772-774.
5. Makino T, Gojobori T: **The evolutionary rate of a protein is influenced by features of the interacting partners.** *Mol Biol Evol* 2006, **23**:784-789.
6. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes.** *Mol Biol Evol* 2005, **22**:598-606.
7. Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY: **Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota.** *J Mol Evol* 2006, **63**:1-11.
8. Yamada T, Bork P: **Evolution of biomolecular networks: lessons from metabolic and protein interactions.** *Nat Rev Mol Cell Biol* 2009, **10**:791-803.
9. Castro MA, Dalmolin RJ, Moreira JC, Mombach JC, de Almeida RM: **Evolutionary origins of human apoptosis and genome-stability gene networks.** *Nucleic Acids Res* 2008, **36**:6269-6283.
10. Barabasi AL, Oltvai ZN: **Network biology: Understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101-U15.
11. Holme P: **Metabolic robustness and network modularity: a model study.** *Plos One* 2011, **6.**
12. Edwards JS, Ibarra RU, Palsson BO: **In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data.** *Nat Biotech* 2001, **19**:125-130.
13. Harrison R, Papp B, Pál C, Oliver SG, Delneri D: **Plasticity of genetic interactions in metabolic networks of yeast.** *PNAS* 2007, **104**:2307-2312.
14. Harrington ED, Jensen LJ, Bork P: **Predicting biological networks from genomic data.** *FEBS Lett* 2008, **582**:1251-1258.
15. Bult CJ, Blake JA, Richardson JE, Kadin JA, Eppig JT, Baldarelli RM, Barsanti K, Baya M, Beal JS, Boddy WJ, Bradt DW, Burkart DL, Butler NE, Campbell J, Corey R, Corbani LE, Cousins S, Dene H, Drabkin HJ, Frazer K, Garippa DM, Glass LH, Goldsmith CW, Grant PL, King BL, Lennon-Pierce M, Lewis J, Lu I, Lutz CM, *et al*: **The Mouse Genome Database (MGD): integrating biology with the genome.** *Nucleic Acids Res* 2004, **32**:D476-D481.
16. Hirschman JE, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Hong EL, Livstone MS, Nash R, Park J, Oughtred R, Skrzypek M, Starr B, Theesfeld CL, Williams J, Andrada R, Binkley G, Dong Q, Lane C, Miyasato S, Sethuraman A, Schroeder M, Thanawala MK, Weng S, Dolinski K, Botstein D, Cherry JM: **Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome.** *Nucl Acids Res* 2006, **34**:D442-D445.
17. Koonin EV: **Orthologs, paralogs, and evolutionary genomics.** *Annual Review of Genetics* 2005, **39**:309-338.
18. Cai JJ, Petrov DA: **Relaxed purifying selection and possibly Hhigh rate of adaptation in primate lineage-specific genes.** *Genome Biol Evol* 2010, **2**:393-409.
19. Vibranovski MD, Zhang Y, Long M: **General gene movement off the X chromosome in the Drosophila genus.** *Genome Res* 2009, **19**:897-903.
20. Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M: **Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome.** *PLoS Biol* 2010, **8**:e1000494.
21. Shannon CE: **A Mathematical theory of communication.** *Bell System Technical Journal* 1948, **27**:379-423.
22. Long M, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nat Rev Genet* 2003, **4**:865-875.
23. Lynch M: **Genomics Gene duplication and evolution.** *Science* 2002, **297**:945-947.
24. Zhang JZ: **Evolution by gene duplication: an update.** *Trends in Ecology & Evolution* 2003, **18**:292-298.
25. Conant GC, Wolfe KH: **Turning a hobby into a job: How duplicated genes find new functions.** *Nat Rev Genet* 2008, **9**:938-950.
26. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.** *Genome Biol* 2004, **5**:R7.
27. Barkman T, Zhang J: **Evidence for escape from adaptive conflict?** *Nature* 2009, **462**:E1-E3.
28. Des Marais DL, Rausher MD: **Escape from adaptive conflict after duplication in an anthocyanin pathway gene.** *Nature* 2008, **454**:762-765.
29. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nat Rev Genet* 2010, **11**:97-108.
30. Miller W, Makova KD, Nekrutenko A, Hardison RC: **Comparative genomics.** *Annu Rev Genomics Hum Genet* 2004, **5**:15-56.
31. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**:1596-1599.
32. Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins.** In *Evolving genes and proteins.* Edited by: Bryson V, Vogel HJ. New York, Academic Press; 1965:97-166.
33. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* .
34. Zhao J, Ding GH, Tao L, Yu H, Yu ZH, Luo JH, Cao ZW, Li YX: **Modular co-evolution of metabolic networks.** *BMC Bioinformatics* 2007, **8.**
35. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M: **Natural selection shapes genome-wide patterns of copy-number polymorphism in drosophila melanogaster.** *Science* 2008, **320**:1629-1631.
36. Glazko G, Mushegian A: **Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns.** *Genome Biology* 2004, **5**:R32.
37. Qian X, Yoon BJ: **Effective identification of conserved pathways in biological networks using hidden Markov models.** *PLoS One* 2009, **4**: e8070.
38. Chen S, Zhang YE, Long M: **New genes in Drosophila quickly become essential.** *Science* 2010, **330**:1682-1685.
39. Jones CD, Begun DJ: **Parallel evolution of chimeric fusion genes.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:11373-11378.
40. Piatigorsky J, Wistow G: **The recruitment of crystallins: new functions precede gene duplication.** *Science* 1991, **252**:1078-1079.
41. Chen FC, Chen CJ, Li WH, Chuang TJ: **Gene family size conservationis a good indicator of evolutionary rates.** *Mol Biol Evol* 2010, **27**:1750-1758.
42. Wilson AC, Carlson SS, White TJ: **Biochemical evolution.** *Annual Review of Biochemistry* 1977, **46**:573-639.
43. Hsiao C, Mohan S, Kalahar BK, Williams LD: **Peeling the onion: ribosomes are ancient molecular fossils.** *Mol Biol Evol* 2009, **26**:2415-2425.
44. Caetano-Anollqs G, Yafremava LS, Gee H, Caetano-Anollqs D, Kim HS, Mittenthal JE: **The origin and evolution of modern metabolism.** *The International Journal of Biochemistry & Cell Biology* 2009, **41**:285-297.
45. Matias Rodrigues JF, Wagner A: **Evolutionary plasticity and innovations in complex metabolic reaction networks.** *PLoS Comput Biol* 2009, **5**: e1000613.
46. Dandekar T, Schuster S, Snel B, Huynen M, Bork P: **Pathway alignment: application to the comparative analysis of glycolytic enzymes.** *Biochem J* 1999, **343**:115-124.
47. Chang X, Wang Z, Hao P, Li YY, Li YX: **Exploring mitochondrial evolution and metabolism organization principles by comparative analysis of metabolic networks.** *Genomics* 2010, **95**:339-344.
48. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of Rickettsia prowazekii and the origin of mitochondria.** *Nature* 1998, 133-140.
49. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucl Acids Res* 2009, **37**:D412-D416.
50. O'Brien KP, Remm M, Sonnhammer ELL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucl Acids Res* 2005, **33**:D476-D480.
51. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucl Acids Res* 1999, **27**:29-34.
52. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.
53. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, Rao BS, Smirnov S, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
54. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA: **The Mouse Genome Database (MGD): mouse biology and model systems.** *Nucleic Acids Res* 2008, **36**:D724-D728.
55. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S: **Genew: the Human Gene Nomenclature Database, 2004 updates.** *Nucl Acids Res* 2004, **32**: D255-D257.
56. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D: **SGD: Saccharomyces Genome Database.** *Nucleic Acids Res* 1998, **26**:73-79.

57. Hooper SD, Bork P: **Medusa: a simple tool for interaction graph analysis.** *Bioinformatics* 2005, **21**:4432-4433.
58. Castro MAA, Filho JLR, Dalmolin RJS, Sinigaglia M, Moreira JCF, Mombach JCM, Almeida RMC: **ViaComplex: software for landscape analysis of gene expression networks in genomic context.** *Bioinformatics* 2009, **25**:1468-1469.

doi:10.1186/1745-6150-6-22
Cite this article as: Dalmolin *et al.*: Evolutionary plasticity determination by orthologous groups distribution. *Biology Direct* 2011 **6**:22.

Capítulo III

Preferential duplication of intermodular hub genes: an evolutionary signature in genome networks.

Artigo científico submetido ao periódico *Plos One*.

Preferential duplication of intermodular hub genes: an evolutionary signature in genome networks.

Ricardo M. Ferreira*[2],José Luiz Rybarczyk-Filho*[2], Rodrigo J. S. Dalmolin*[3], Mauro A. A. Castro[1,2], José C. F. Moreira[3], Leonardo G. Brunnet[2]& Rita M. C. de Almeida[1,2]

National Institute of Science and Technology for Complex Systems[1], Instituto de Física[2], and Departamento de Bioquímica[3], Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, 91051-970 C.P. 15051, Porto Alegre, Brazil.

***These authors contributed equally to this paper**

Correspondence to:

Rita M. C. de Almeida
Instituto de Física, Universidade Federal do Rio Grande do Sul,
Av. Bento Gonçalves, 9500, 91051-970 C.P. 15051, Porto Alegre, Brazil.

Running head:
An evolutionary signature in genome networks

Keywords: Gene network, Genome evolution, Protein-protein interaction matrix, Monte Carlo dynamics.

ABSTRACT

Whole genome protein-protein association networks are not random and their topological properties stem from genome evolution mechanisms. In fact, more connected, but less clustered proteins are related to genes that, in general, present more paralogs as compared to other genes, indicating frequent previous gene duplication episodes. On the other hand, genes related to conserved biological functions present few or none paralogs and yield proteins that are highly connected and clustered. These general network characteristics must have an evolutionary explanation. Considering data from STRING database, we present here experimental evidence that, more than not being scale free, protein degree distributions of organisms present an increased probability for high degree nodes. Furthermore, based on this experimental evidence, we propose a simulation model for genome evolution, where genes in a network are either acquired *de novo* using a preferential attachment rule, or duplicated with a probability that linearly grows with gene degree and decreases with its clustering coefficient. For the first time a model yields results that simultaneously describe different topological distributions. Also, this model correctly predicts that, to produce protein-protein association networks with number of links and number of nodes in the observed range, it is necessary 90% of gene duplication and 10% of *de novo* gene acquisition.

This scenario implies a universal mechanism for genome evolution.

INTRODUCTION

Genome evolution is determined first by the processes that modify DNA and then by those mechanisms that either neutrally keep or naturally select these mutations by their phenotypic effects. The connection between DNA variations and the consequent phenotypic alterations is far from being simple and is elusive to determine. However, it is reasonable to assume that, after evolutionary time spans, these DNA variation mechanisms have left their mark on the genome.

Phenotypic effects are consequence of the existing associations between proteins which rule cellular metabolism. As proteins are expressed from genes, protein-protein associations will express eventual changes in genotypes and are prone to natural selection. Consequently we may speculate that natural selection, by defining genome evolution mechanisms, has left its mark on organisms' protein-protein association matrices. This is not a novel idea. Barabási and collaborators [1,2] have described genomes of different organisms as networks where nodes are either genes or proteins, and links correspond to associations between the nodes. They proposed an evolution dynamics for the genome considering that genes are sequentially added to a network following a preferential attachment rule: each newly incorporated gene interacts with a gene already on the network with a probability that is proportional to its degree, that is, to the number of other genes with which it already interacts. The resulting artificial network is scale free and described well the available experimental data at that date.

However, the properties of a gene already in the network are not the only drive

for a novel gene attachment. There are different molecular mechanisms acting as novelty source in gene formation, such as exon shuffling, retroposition, mobile elements, horizontal gene transfer, gene duplication, etc., and the connections of a new gene certainly reflect its origin together with the nature of the genes it connects to [3]. Among the mechanisms involved in new genes creation, gene duplication is recognizably the most important and there is plenty of evidence that it plays an essential role on genome evolution [4]. One major feature of a duplicated gene consists of inheriting its parent connections and this property is determinant to the whole network design.

Vázquez and collaborators [5,6] proposed a model for genome evolution where genes are incorporated by duplication followed by mutations which are translated as adding and/or deleting links on a protein-protein association matrix. In this model, genes are randomly chosen to duplicate and parameters are set to produce gene networks where the probability that a gene product is associated to k other proteins decays as a power law as k increases. A drawback for this approach, using randomly chosen genes, lays on the experimental fact that the probability to fix a given duplication episode greatly varies according to the properties of the duplicating gene [7–9].

Since the contributions by Barabási and collaborators, the amount and quality of data regarding both genomes and protein-protein association have greatly increased. For example, STRING database increased from few organisms at 2001 to 1133 organisms in 2011 [10–12]. Also, databases regarding protein-protein association for some organisms have been largely enhanced. Here we analyze data considering 268 core organisms, which strongly suggest that highly connected genes stem from duplication mechanisms acting preferentially on genes that are highly connected, but

not excessively clustered. These conclusions are made evident here by presenting the quantities as functions of $\frac{k}{k_{max}}$, where k_{max} is the maximum degree in the network. We also propose an adequate ordering for genes to globally evince topological properties of the protein-protein association matrix.

Considering these experimentally based conclusions we propose a genome evolution dynamics where the probability that a gene duplicates grows with its degree and decreases depending on how clustered it is. We also consider a Barabási mechanism of acquiring genes *de novo* based on preferential attachment. The results of these simulations are capable of describing different aspects of the network topology, besides predicting the ratio of duplicated and *de novo* acquired genes.

RESULTS

Building protein-protein association matrices. We considered all 268 core organisms in STRING database, version 8.3 [10–12], with confidence scores 0.700, 0.800, and 0.900 using "experimental" and "database" (95% of these interactions) added with "neighborhood", "fusion", "co-expression", and "co-occurrence" evidence. This information renders possible to build a network, where each node corresponds to a protein with at least one known protein-protein association, and links correspond to these associations. To each network node i we assign a degree k_i, which is the number of links arriving at that node. For each organism and score we produce a network and calculate the probability $P'(k)$ that a protein has k links, defined as

$$P'(k) = \frac{N(k)}{N} \quad (1)$$

where N is the number of nodes and $N(k)$ is the number of nodes with degree k. To compare different organisms, with different genome sizes, we considered a rescaled probability of finding a protein with a given degree k, as follows

$$P\left(\frac{k}{k_{\max}}\right) = k_{\max} P'(k) \quad (2)$$

Where $k_{\max}$ is the maximum degree presented by the proteins of an organism.

Figure 1a presents the average, taken in intervals of $\frac{\Delta k}{k_{\max}} = 0.01$, of the network degree distribution, $P\left(\frac{k}{k_{\max}}\right)$ versus for three different confidence scores: 0.700, 0.800 and 0.900. The inset presents the degree distributions of all 268 core organisms, with different colors for different scores. The blue line in Fig. 1a is a power law fit, $F\left(\frac{k}{k_{\max}}\right) = 0.02\left(\frac{k}{k_{\max}}\right)^{-2.4}$, which describes $P\left(\frac{k}{k_{\max}}\right)$ for only a limited interval of $\frac{k}{k_{\max}}$. At values of $\frac{k}{k_{\max}}$ near 0.9, this degree distribution presents a local maximum, associated to the cloud of points with higher values of probability presented in the inset. The probability of proteins with degree near $k_{\max}$ increases and indicates a genome evolution dynamics where high degree genes are probable to appear. As the main mechanism of genome evolution is gene duplication [3,4], it is reasonable to assume that the local maximum in $P\left(\frac{k}{k_{\max}}\right)$ for large $\frac{k}{k_{\max}}$ is due to high duplication probability for more connected genes. Figure 1b presents the same

data in a linear plot, where the standard deviations for each average value of $P\left(\frac{k}{k_{\max}}\right)$ are shown, to evince that deviations from the power law fit is significant. Each point is an average over 268 organisms, justifying a Z test for significance. The difference between the power law fit and the average $P\left(\frac{k}{k_{\max}}\right)$ for confidence score 0.800 is shown in the inset for Fig.1b, in units of standard deviations for $P\left(\frac{k}{k_{\max}}\right)$, calculated in intervals of $\frac{\Delta k}{k_{\max}} = 0.01$. The maximum in degree distribution is significantly different from the power law. This is a novel result which has been evinced by plotting the distributions as functions of $\frac{k}{k_{\max}}$, instead of functions of k or $\frac{k}{N}$. From now on, we shall refer to $\frac{k}{k_{\max}}$ as the relative degree of a node, which varies in the interval (0,1).

Figure 1c plots as a function of $\frac{k}{k_{\max}}$ the average clustering coefficient $\langle C \rangle$, defined as the fraction of existing connections between the neighbors of a gene with k neighbors in relation to the maximum number of such connections $\frac{k(k-1)}{2}$. The inset in Fig. 1c individually shows the corresponding data for all core organisms. For all three scores this curve is initially constant, presenting local minimum and maximum for, roughly, $\frac{k}{k_{\max}} \approx 0.02$ and $\frac{k}{k_{\max}} \approx 0.8$, respectively, decreasing after that: the most connected genes are not the maximally clustered. Observe that, while the maximum in $P\left(\frac{k}{k_{\max}}\right)$ occurs for $\frac{k}{k_{\max}} \approx 0.9$, the maximum for the clustering

coefficient occurs before that.

Figure 1d plots the average relative degree of the neighbors $\langle k_{nn} \rangle$ of a gene as a function of $\frac{k}{k_{\max}}$. The inset individually shows the corresponding data for all core organisms. For all scores this curve is initially increasing, presenting a local maximum at roughly $\frac{k}{k_{\max}} \approx 0.9$, decreasing after that. It means that the most connected genes are not connected to the highest k genes. Observe also that the maxima in both $P\left(\frac{k}{k_{\max}}\right)$ and $\langle k_{nn} \rangle$ occur for $\frac{k}{k_{\max}} \approx 0.9$.

Summarizing, these plots indicate that *i)* $P\left(\frac{k}{k_{\max}}\right)$ is not power law; *ii)* $P\left(\frac{k}{k_{\max}}\right)$ presents a local maximum for $\frac{k}{k_{\max}} \approx 0.9$; *iii)* the clustering coefficient is not uniform, presenting local minimum and maximum; and *iv)* the network is assortative up to $\frac{k}{k_{\max}} \approx 0.9$, with $\langle k_{nn} \rangle$ decreasing after that. These observations suggest modules of high average degree which are highly clustered. This behavior is evinced by the superposition of data from a large number of organisms, plotted against a normalized degree $\frac{k}{k_{\max}}$. For comparison, Fig. S1 presents plots where the degree k is normalized by the total number of genes of each organism: this behavior is not as clearly unveiled.

Another experimental aspect is relevant for genome evolution. Duplication events can be assessed by analyzing gene families, *i.e.*, genes sharing the same ancestral gene. Some gene families have mainly orthologs, while others are

composed by a great number of paralogs, indicating many duplication episodes [7,13]. The reason why some genes are prone to duplicate while others avoid duplication is controversial. However, duplication is clearly not randomly fixed and functional characteristics of the parent gene certainly influence new born genes fates. It has been discussed that genes presenting substrate promiscuity are prone to fix duplication while other genes avoid duplication because it probably leads to deleterious effects [14].

Li and collaborators [15] demonstrated that highly connected proteins with low clustering coefficient (intermodular hubs) possess a higher proportion of duplicated genes as compared with proteins that are highly connected and highly clustered (intramodular hubs). According to those authors, intramodular hubs represent the network most stable and conservative part, while intermodular hubs represent evolutionary dynamic network regions with a high duplication rate. Similar results has been found by Fraser [16].

Genome evolution model. These experimentally determined characteristics of genomes may be explained by an evolution dynamics with two different gene acquisition mechanisms: *de novo* formation and duplication. The first mechanism follows Barabási preferential attachment rule, which simulates an enhanced attachment probability shown by genes with more active domains. The second mechanism describes the experimental facts discussed above: genes are chosen with higher probability when they are more connected, but less clustered. Protein-protein association information may be organized as a binary matrix whose elements are noted by M_{ij}, such that $M_{ij} = 1$ in case proteins labeled by indices i and j are associated and $M_{ij} = 0$ otherwise. Now, the clustering coefficient C_i for the i^{th} gene is defined as [17,18]

$$C_i = \frac{2}{k_i(k_i - 1)} \sum_{j=1}^{N} \sum_{l=1}^{N} M_{ij} M_{jl} M_{li} \quad , \tag{4}$$

which gives the ratio of existing links between the neighbors of the i^{th} gene to the maximum possible number of such links (which is equal to the number of combinations of k_i elements 2 by 2).

The duplication probability for the i^{th} gene is defined as

$$p_i^D = \frac{k_i(1 - C_i)}{\sum_j k_j(1 - C_j)} \quad , \tag{5}$$

where the denominator guarantees a normalized probability. This assumption reproduces the experimental facts that *i)* degree distributions have a local maximum for $\frac{k}{k_{\max}}$ near 1 (Fig.1) and *ii)* more clustered genes are less prone to duplicate [7,15,19].

Simulations start with 5 nodes, each linked to two others, forming a ring. To acquire a new gene we first choose either *de novo* mechanism, with probability $(1-q)$, or duplication, with probability q. If the *de novo* mechanism is chosen, each existing node i is linked to the new one with probability $\frac{k_i}{\sum_j k_j}$, and the procedure is repeated until the new node presents at least one link. In case of duplication, the node to be duplicated is chosen by using the probability defined in Eq.(5). Duplication implies creating a new node linked to its parent and with the same neighbors.

After duplication, mutations are implemented by deleting links between either the parent or the child with a common neighbor with probability r. In fact, a

hallmark of gene duplication is the subsequent speciation of at least one gene copy [20].

To compare with simulated genome evolution dynamics we chose those organisms for which there is more information regarding protein-protein association. Figure 2a shows the number of links versus the number of genes for the 268 core organisms for 0.800 confidence score. Observe that data for very well studied organisms as *Homo sapiens* or *Arabidopsis thaliana*, present larger numbers of genes and links, that is, more information is available. In what follows we considered 6 organisms, marked with orange dots in Fig. 2a (*Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Escherichia coli).*

The present simulation model has two parameters, duplication probability q and mutation probability r. For the numbers of links and genes of simulated networks to fall in the same intervals as more extensively investigated organisms (Fig. 2a), q must be of the order of 0.90, which is experimentally verified: Zhou *et al.*[4] have studied *Drosophila melanogaster* genome and compared it to other organisms in *D. melanogaster* subgroup. They have found that duplication is responsible for 80% of new genes, and 10% is generated by retroposition, here taken as an additional form of gene duplication. We are left with one single parameter r, set to 0.05 to match the observed relation between number of links and nodes presented by protein-protein association matrices of real organisms (Fig.2a).

We also simulated two other well described models for genome evolution: Barabási and Albert model [1], based on a preferential attachment rule, and Vazquez *et al.* [5,6] model, where genomes are built by duplicating randomly chosen genes. For both models, parameters are set to ensure that the number of links and nodes are

roughly the same as in the protein-protein association networks obtained from STRING database for confidence score 0.800. In Barabási-Albert model, each new node is connected with 15 neighbors, and in the duplication-divergence model each node is linked with its parent, and has 0.4 of mutation probability. For brevity, we considered the most cited models in the literature although other interesting models also address genome growth [21–24].

Figs. 2a, 2c, and 2e present, as a function of N, the plots of number of links N_L, average degree $\langle k \rangle$, and maximum degree, $k_{\max}$, for experimental results (dots) and simulated models (solid lines). As discussed, the chosen model parameters ensure that the simulated number of links crosses the region with best investigated organisms. The experimental points indicate that the number of links is proportional to the number of nodes, that is, $N_L \sim N^1$. This behavior is clearly shown by both Barabási-Albert and our model, and is further evinced by Fig. 2c, that shows a constant average degree for experimental dots and these two models. Finally, Fig. 2e shows that, for the simulations, $k_{\max}$ increases with, roughly, $\sqrt{N}$. The experimental results are not in contradiction, although they are not conclusive. Anyway, this behavior explains why using $k_{\max}$ instead of N as the normalization constant in Eq. 2 yield different results.

Figs. 2b, 2d, and 2f present $P\left(\frac{k}{k_{\max}}\right)$ versus $\frac{k}{k_{\max}}$ for the three simulated models, measured at different instants. Observe that clearly Barabási-Albert and our model converge to a scaling invariant distributions that superpose as $N \to \infty$, while for the Vázquez (D-D) model this convergence is either not true or too slow. This is a relevant point: although real genomes are finite, we may speculate that when large

enough they present a scale invariant degree distribution. If this is true, the data collapse predicted by scaling invariance, together with a significant fit of the collapsed degree distribution of all core organisms, is as a strong evidence of a common mechanism universally ruling genome growth.

On the other hand, experimental degree distributions may present finite size effects. This is clear in Fig. S2, where we show $P\left(\frac{k}{k_{\max}}\right)$ versus $\frac{k}{k_{\max}}$ for the experimental data (score 0.800) averaged over genomes whose protein-protein association networks present N in the ranges $N<1000$, $1000<N<2000$, …, $6000<N$. The degree distributions seem to converge to a scale invariant state, but for the smaller networks the finite size effects are visible. Both experimental data and D-A model results show that smaller networks present a higher local maximum in $P\left(\frac{k}{k_{\max}}\right)$ for large $\frac{k}{k_{\max}}$. To properly compare the simulations results with experimental networks with variable sizes, we considered a weighted average of the degree distribution, as follows.

For each model, we produced 10 samples in each size range listed above, and obtained the distributions of degree, clustering coefficient, and average degree of the neighbors as functions of $\frac{k}{k_{\max}}$. To compare with the set of all 268 core organisms, presenting, respectively, 32, 110, 74, 39, 10, 1 and 2 organisms in each size range, we produced weighted averages over the size ranges for the topological distributions, using the weights 32/268, 110/268, 74/268, 39/268, 10/268, 1/268 and 2/268. These results are shown in Fig. 3. Other parameters values in each model yield different results, shown in Figs. S3-S6 in Supplementary Materials: the

description of topological quantities are worse in these cases. Similar averages for the six, best investigated organisms are shown in Fig. S7 of Supplementary Materials.

Duplication-Acquisition model reproduces the topology of protein-protein association networks. For each network, we calculated the weighted average for probability $P\left(\frac{k}{k_{\max}}\right)$, the clustering coefficient $\langle C\rangle_{k/k_{\max}}$, and the relative degree $\langle k_{nn}\rangle_{k/k_{\max}}$ of the neighbors of a node with degree $\frac{k}{k_{\max}}$, defined as

$$\langle k_{nn}\rangle_{k/k_{\max}} = \frac{1}{N(k)}\sum_{i=1}^{N}\frac{\delta(k_i-k)}{k_i}\sum_{\langle j\rangle_i}^{k_i}\frac{k_j}{k_{\max}}, \qquad (6)$$

where $\langle j\rangle_i$ stands for a sum over the nodes j that are neighbors to node i, and $\delta(k_i-k)=1$ if $k_i-k=0$ and $\delta(k_i-k)=0$ otherwise.

The black dots in Figs. 3 refer to protein-protein association networks of the 268 core organisms, which present large clustering coefficients for all degrees, decreasing as $\frac{k}{k_{\max}}$ approaches 1: very high degree nodes are less clustered than less connected nodes. In organisms, the average number of connections of the neighbors, $\langle k_{nn}\rangle$, first increases with the node degree and then decreases, reinforcing the fact of very high degree nodes not presenting the largest clustering coefficient. Figure 3 presents three columns, one for each model, where we show the *i)* the experimental data as black points, weighted averages for *ii)* experimental points as green lines and for *iii)* simulation as red lines. The first column shows that

B-A model produces a degree distribution $P\left(\frac{k}{k_{\max}}\right)$ that follows a power law, a clustering coefficient that is roughly constant at a value that is much less than those shown by experimental data. Furthermore, $\langle k_{nn} \rangle$ does not depend on $\frac{k}{k_{\max}}$. The deviation from the experimental dots reflects that Barabási-Albert model yield scale free networks with a global central hub.

The second column presents the results for the Duplication-Divergence (D-D) model. Here, this distribution clearly does not follow a power law, due to the chosen parameters (link deleting probability of 0.4), that fixed the ratio of number of links to number of nodes to the desired values (see Fig. 2a). The average clustering coefficient decreases too abruptly, as compared to experimental data: as degree increases, the clustering decreases as $\sim\left(\frac{k}{k_{\max}}\right)^{-0.7}$. However, the average degree of the neighbors presents a mild increase, meaning that genes connect to groups of genes with slightly larger degrees.

The third column in Fig. 3 refers to the results of our model. In Fig. 3c, $P\left(\frac{k}{k_{\max}}\right)$ describes very well the experimental data. For high values of $\frac{k}{k_{\max}}$, degree distribution reproduces the local maximum as shown by real organisms, although for smaller degrees. The clustering coefficient, shown in Fig. 3f, describes the major part of the interval, presenting a more intense decrease as $\frac{k}{k_{\max}} \to 1$. The varying character of assortativeness as $\frac{k}{k_{\max}}$ increases is also evident in Fig. 3i: $\langle k_{nn} \rangle$ first increases to a maximum up to $0.45 k_{\max}$.

Comparing the three columns we conclude that D-A model better catches the topological properties of protein-protein association networks, according to the currently available experimental data, although the description is not perfect.

Ordering the protein-protein association matrix evince global network properties.

Furthermore, to evince global properties of the networks, the protein-protein association data that is organized on the matrix M where each axis represents the protein list in a given order. The matrix elements M_{ij} are assigned with value 1 (0) if there is (not) an association between the genes at positions i and j of the list. For illustrational purposes, these association matrices may be represented by plots where a black dot at position (i, j) indicates that $M_{ij} = 1$.

We obtain the sets of genes of each organism from STRING database and dispose them in randomly ordered lists. Each possible order for a gene list implies a different configuration for matrix M, for which a cost function E may be defined as

$$E = \sum_{i=1}^{N} \sum_{j \neq i}^{N} d_{ij}^{\alpha} \left(\left| M_{i,j} - M_{i+1,j} \right| + \left| M_{i,j} - M_{i-1,j} \right| + \left| M_{i,j} - M_{i,j+1} \right| + \left| M_{i,j} - M_{i,j-1} \right| \right), \quad (3)$$

Where $d_{ij} = \sqrt{\left| i^2 - j^2 \right|}$ is proportional to the distance on the matrix from the point $(i.j)$ to the diagonal (when $i = j$), and α is a parameter, here taken $\alpha = 8$. Minimization of this function, by changing the genes localization on the list, implies approximating mutually interacting genes, as discussed by Rybarczyk-Filho *et al*.

[25].

The ordering algorithm starts from a randomly ordered matrix configuration and proceeds by randomly choosing a pair of genes whose positions are tentatively swapped. The cost function for this changed configuration is calculated and, in case the cost decreases, the change is accepted. If the cost function increases by ΔE, the change is accepted with probability $e^{-\Delta E/T}$, where T is a parameter. This procedure is intended to avoid metastable states in the optimization of Eq.(3). Finally, when $\Delta E = 0$, the change is accepted with 50% probability. The algorithm proceeds by randomly choosing another pair of genes and the procedure is repeated until the value of the cost function is stabilized.

Randomly ordered lists yield association matrix configurations with black dots spread over the whole plot. Ordering the gene list by minimizing the cost function evinces topological properties of protein-protein association networks. Figure 4a-f presents the ordered matrices for the six organisms listed above. Observe that points concentrate near the diagonal, implying that there may be an association ($M_{ij} = 1$) between the products of genes localized at not far apart positions i and j. Not all networks may be put in formats like those shown by Figs. 4a-f. See Fig. 4-g which represents a network built using Barabási-Albert algorithm, or an Erdös-Rényi network, presented on Fig. S8 of Supplementary Materials. In fact, this format reveals that genomes (Figs.4a-f) do not present one central hub linked to the whole network (which could indicate scale free networks) but, contrarily, present many hubs with neighborhoods that do not span the entire system.

Figures 4g-i present ordered association matrices for simulated networks. Barabási-Albert (B-A) model (Fig. 4g) clearly shows only one module, with a

central hub connected to all network. Duplication-Divergence (D-D) model, on the other hand, shows a slimmer structure around the diagonal, and Duplication-Acquisition (D-A) model presents a central hub not connected to the whole network. Figure S9 of Supplementary Materials presents the same panels, zooming at the central regions: the hierarchical structure of clusters, evinced by small solid squares, is clearly present in organisms and Duplication-Acquisition model. Figure S10 of Supplementary Materials present the orderings obtained with $\alpha = 1$, which stresses further the clustered structures.

Together, figures 1 and 4 evince different aspects of real genomes. First, degree distribution is not a power law. Second, there is an accumulation of high degree nodes, which may be explained by an enhanced duplication probability for highly connected gene products. Finally, hub genes are not central to the whole network, which presents hierarchical clusters.

DISCUSSION AND CONCLUSIONS

In this paper we have presented experimental evidence that degree distribution is not scale free, presenting an increased probability for high degree nodes, and that there are a few hub nodes in these networks, probably organized in a hierarchical way. Furthermore, when scaled by the maximum degree in each network, $k_{\max}$, the degree distribution seems to approach a scale invariant state as the number of genes in the network increases. However, real genomes still present finite size effects. This scenario indicates a universal mechanism for genome evolution.

The understanding of genome growth mechanisms is a central point in

evolutionary biology. It is well established that gene duplication is the main process for new genes emergence. Therefore, it is reasonable to think that gene duplication represents an essential feature for genome evolution. This idea has been used by Vázquez in his genome evolution model including gene duplication as genetic novelty source [5,6]. However, in that model genes are randomly chosen to duplicate whereas experimental evidence indicates that gene duplication is not random. There are huge differences in the fixation probability of a gene duplication event. Depending on gene niche, the new copy could be selectively fixed or eliminated [14]. This concept becomes clear when gene families are assessed. There are some gene families composed basically by vertically inherence (*i.e.* orthologs), without duplication episodes. On the other hand, there are gene families composed by great number of duplication-generated genes (*i.e.* paralogs) [7,13]. The question is what gene characteristics will increase the fixation probability of its duplication?

The local maximum shown in Figure 1a gives us a clue about gene duplication dynamics. According to the figure, there is an increased probability of very connected proteins, indicating a genome evolution dynamics favoring hub genes emergence. However, there are at least two very distinct classes of hub genes: (*i)* intramodular hubs, presenting high degree and high clustering coefficient, and (*ii*) intermodular hubs, presenting high degree and low clustering coefficient. The first one takes part in modules, which generally comprises intricate biological systems where all proteins exercise coordinate functions. In many of those systems, stoichiometry relationship is needed and a duplication event could be deleterious to the whole system. The second connects different modules, commonly exercising pleiotropic functions. Gene duplication theories always associate the fixation of the new-born gene copy whit new function development [20]. Additionally, a gene

performing more than one function - when each function cannot be independently optimized - could benefit from a duplication event where each gene copy is rendered free to independently optimize different functions [26].

Intermodular hubs have been discussed as targets of gene duplication [15]. Also, Szklarczyk et al have shown that for yeast in nearly 70% of small scale duplication events, the paralogs do not remain working in the same complex and in at least 40% their ancestor gene should participate in more than one biological module [27]. On the other hand, intramodular hubs are associated to ancient networks that have reached their architecture early in evolution and any modification can affect their homeostasis [7]. This fact is well exemplified by ribosomes and DNA repair mechanisms, both very ancient systems with modular network architecture and both composed by genes with almost none duplication episode fixed though their evolutionary history [7,19] .

Here, we proposed a simulation model for genome evolution, Duplication-Acquisition model, where genes in a network are either duplicated or acquired *de novo* using a preferential attachment rule. However, according to our model, genes are not arbitrarily chosen to duplicate: the duplication probability linearly grows with gene degree and decreases with its clustering coefficient. In other words, intermodular hubs have increased probability to duplicate. With this simple rule, topological distributions of biological networks are well described. This model correctly predicts that, to produce protein-protein association networks with number of links and number of nodes in the observed range, it is necessary 90% of gene duplication and 10% of *de novo* gene acquisition.

To compare the networks we ordered gene lists for each organism and model to

produce protein-protein association matrices yielding images of the network association structure. These images give a global assessment of the networks, suggesting that there is a system scale that is less than its size (see Fig.4), with, possibly, a hierarchical modular organization, as predicted by the Duplication-Acquisition model (see Fig. S10).

The simulation model is not perfect. Most probably phenotypic effects caused by gene acquisition, duplication, or mutation cannot be fully grasped by network gene properties only and, consequently, this model is an over-simplification. However it does point towards a positive correlation between duplication probability and degree, while indicating a negative correlation between duplication probability and clustering coefficient. Consequently, Duplication-Acquisition model suggest how and where evolution works to build genetic novelty.

ACKNOWLEDGEMENTS

We acknowledge fruitful discussions with Prof. Diego Bonnato, Centro de Biotecnologia, and support from the Centro de Física Computacional, Universidade Federal do Rio Grande do Sul.

FUNDING

This work has been partially funded by Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

Reference List

1. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512. 7898 [pii].

2. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407: 651-654. 10.1038/35036627 [doi].

3. Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4: 865-875.

4. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W (2008) On the origin of new genes in Drosophila. Genome Res 18: 1446-1455. gr.076588.108 [pii];10.1101/gr.076588.108 [doi].

5. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of Protein Interaction Networks. Complexus 1: 38-44.

6. Vázquez A (2003) Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. Phys Rev E Stat Nonlin Soft Matter Phys 67: 056104.

7. Dalmolin R, Castro M, Rybarczyk Filho J, Souza L, de Almeida R, Moreira J (2011) Evolutionary plasticity determination by orthologous groups distribution. Biology Direct 6: 22.

8. Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet 11: 487-498. nrg2810 [pii];10.1038/nrg2810 [doi].

9. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483-5488. 0501761102 [pii];10.1073/pnas.0501761102 [doi].

10. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucl Acids Res 37: D412-D416.

11. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucl Acids Res 33: D433-D437.

12. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucl Acids Res 35: D358-D362.

13. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5: R7.

14. Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet 9: 938-950.

15. Li L, Huang Y, Xia X, Sun Z (2006) Preferential Duplication in the Sparse Part of Yeast Protein Interaction Network. Mol Biol Evol 23: 2467-2473.

16. Fraser HB (2005) Modularity and evolutionary constraint on proteins. Nat Genet 37: 351-352.

17. Colizza V, Flammini A, Maritan A, Vespignani A (2005) Characterization and modeling of proteinGÇôprotein interaction networks. Physica A: Statistical Mechanics and its Applications 352: 1-27.

18. Costa LD, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: A survey of measurements. Advances in Physics 56: 167-242.

19. Castro MA, Dalmolin RJ, Moreira JC, Mombach JC, de Almeida RM (2008) Evolutionary origins of human apoptosis and genome-stability gene networks. Nucleic Acids Res 36: 6269-6283.

20. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97-108.

21. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4: 51. 1471-2148-4-51 [pii];10.1186/1471-2148-4-51 [doi].

22. Evlampiev K, Isambert H (2007) Modeling protein network evolution under genome duplication and domain shuffling. BMC Syst Biol 1: 49. 1752-0509-1-49 [pii];10.1186/1752-0509-1-49 [doi].

23. Kim WK, Marcotte EM (2008) Age-Dependent Evolution of the Yeast Protein Interaction Network Suggests a Limited Role of Gene Duplication and Divergence. PLoS Comput Biol 4: e1000232.

24. Takemoto K, Oosawa C (2007) Modeling for evolving biological networks with scale-free connectivity, hierarchical modularity, and disassortativity. Math Biosci 208: 454-468. S0025-5564(06)00224-0 [pii];10.1016/j.mbs.2006.11.002 [doi].

25. Rybarczyk-Filho JL, Castro MA, Dalmolin RJ, Moreira JC, Brunnet LG, de Almeida RM (2010) Towards a genome-wide transcriptogram: the Saccharomyces cerevisiae case. Nucleic Acids Res . gkq1269 [pii];10.1093/nar/gkq1269 [doi].

26. Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454: 762-765.

27. Szklarczyk R, Huynen M, Snel B (2008) Complex fate of paralogs. Bmc Evolutionary Biology 8: 337.

FIGURE LEGENDS

Figure 1 – Topological quantities for all 268 core organisms from STRING database. Three different confidence scores: 0.700, 0.800 and 0.900 (black, red and green lines in all graphs, respectively). All measurements are taken as functions of node degree, rescaled by the maximum degree of the corresponding network. All averages were taken over intervals $\Delta k / k_{\max} = 0.01$. (**a**) Average degree distribution compared with a tentative power law fit (blue line). (**b**) Average degree distribution in linear scale, showing the increase in the degree distribution for higher degree. The inset presents the distance between the power law fit and the average of networks with score 0.800 measured in number of standard deviations. (**c)** Clustering coefficient and (**d)** mean nearest neighbor degree averaged over all core organisms. The insets in panels (**a**), (**c**) and (**d**) show individual results for all core organisms for each score.

Figure 2 - Evolution of simulated models. Barabási-Albert, duplication-divergence and duplication-acquisiton networks (red, blue and green lines, respectively). The black dots represent all core organisms from STRING database, where six well studied organisms are highlighted in orange. (**a**) Number of links, (**c**) mean degree and (**e**) maximum degree are shown as functions of the total number of nodes in the network. The degree distribution was calculated in five snapshots of

the evolution of (**b**) Barabási-Albert, (**d**) duplication-divergence, and (**f**) duplication-acquisition models, in intervals of 2000 nodes.

Figure 3 - **Comparison of topological measures for simulated networks.** The black dots represent the superposed networks for all core organisms from string database with confidence score 0.800, the green lines are averages taken in intervals of $\Delta k / k_{\max} = 0.01$, and the red lines are weighted averages of simulated networks. The upper, central, and lower rows show, respectively, degree distribution, clustering coefficient, and nearest neighbor mean degree. Each column refers to a simulated model: Barabási-Albert on the left, duplication-divergence on the center and duplication-acquisition on the right.

Figure 4 - **Ordered association matrices.** This figure presents the association matrices for *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Escherichia coli*, Barbási-Albert model, duplication-divergence model and duplication-acquisition model after running the ordering algorithm. The black dots represent interactions between two nodes.

Supporting Information Figures - Legends

Figure S1. Topological measures for all core organisms. STRING database with confidence scores of 0.700, 0.800, and 0.900 (black, red and green dots), with degree rescaled by number of nodes *N*. Figure (**a**) shows degree distribution, (**b**) clustering coefficient, (**c**) number of links per number of nodes, and (**d**) mean average degree of nearest neighbors. In these figures we can see that the properties discussed in the main text are not clearly evinced.

Figure S2. Degree distribution of protein-protein association matrices relative to core organisms for STRING confidence score 0.800, averaged over intervals $\frac{\Delta k}{k_{\max}} = 0.01$. Each line corresponds to a network in a different range of number of nodes N, as described in the legend.

Figure S3. Four networks obtained using Barabási-Albert model. Different values of parameter *m*, which determines the number of links of each new node. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. In Figure (**a**) we can see that the degree distribution follows a power-law and does not correctly represent the degree distribution of the organisms networks. Figure (**b**) shows that clustering coefficient of the simulated networks is lower than the experimental data. Figure (**c**) presents the evolution of number of links with number of nodes. Figure (**d**) shows that the average degree of the neighbors of a node is independent of the node degree for networks built using Barabási-Albert

model, what deviates from the behavior presented by the organisms networks that are highly assortative.

Figure S4. Three networks obtained using Duplication-Divergence model. Different values of parameter *p*, which determines the mutation probability. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. In Figure (**a**) we can see that for higher values of *p* the network approaches a power-law, but as we can see in Figure (**c**), the number of links fall below those found for the organisms. Figure (**b**) shows that the clustering decreases with degree. In Figure (**d**) we have the average degree of nearest neighbors, which increases with degree, showing that the Duplication Divergence model builds networks with the same assortativeness of the organisms.

Figure S5. Five networks obtained using the Duplication-Acquisition model. Different values of parameter *q*, which determines the fraction of nodes acquired by duplication, maintaining constant *r*, the mutation probability. The grey dots represent the networks for all 268 core organisms, with confidence score 0.800. We can see that, as the number of acquired nodes increases, the network approaches a Barabási-Albert one, as we can see in Figures (**a**), (**b**), and (**d**). Namely, the network loses the high probability of finding high degree nodes in the degree distribution (Figure (**a**)), the clustering coefficient decreases (Figure (**b**)), and the network loses its assortativity (Figure (**d**)). Figure (**c**) shows that the number of links also decreases, falling below the value presented by organisms.

Figure S6. Seven networks obtained using the Duplication-Acquisition model. **D**ifferent values of parameter *r*, which determines the mutation probability, maintaining constant *r*, the fraction of nodes acquired by duplication. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. We can see that, as mutation probability increases, the network approaches the ones obtained using the Duplication-Divergence model. In Figure (**a**) the degree distribution approaches a power-law, and in (**b**) the clustering coefficient decreases.

Figure (**c**) shows that the number of links decreases, and in Figure (**d**) we can see the mean nearest degree distribution for the networks.

Figure S7. Comparison of topological measures for simulated networks. The black dots represent the superposed networks for six organisms from STRING database with confidence score 0.800 (*Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Escherichia coli)*, the red lines are averages of these networks taken in intervals $\frac{\Delta k}{k_{\max}} = 0.01$, and the green lines are weighted averages of simulated networks. Upper, central, and lower rows show, respectively, degree distribution, clustering coefficient, and nearest neighbor mean degree. Each column refers to a simulated model: Barabási-Albert on the left, duplication-divergence on the center and duplication-acquisition on the right.

Figure S8. Association matrix for a Erdös-Rényi network . 4665 nodes and 94830 links, ordered using α=1.

Figure S9. Zoom at the central part of association matrices in Fig.3. From 0.4N to 0.6 N, for **(a)** *Homo sapiens*, **(b)** *Mus musculus*, **(c)** *Arabidopsis thaliana*, **(d)** *Drosophila melanogaster*, **(e)** *Saccharomyces cerevisiae*, **(f)** *Escherichia coli*, **(g)** Barabási-Albert model, **(h)** duplication-divergence model and **(i)** duplication-acquisition model, ordered using α=8.

Figure S10. Association matrices. r **(a)** *Homo sapiens*, **(b)** *Mus musculus*, **(c)** *Arabidopsis thaliana*, **(d)** *Drosophila melanogaster*, **(e)** *Saccharomyces cerevisiae*, **(f)** *Escherichia coli*, **(g)** Barabási-Albert model, **(h)** duplication-divergence model and **(i)** duplication-acquisition model, ordered using α=1.

Figure 1
Click here to download high resolution image

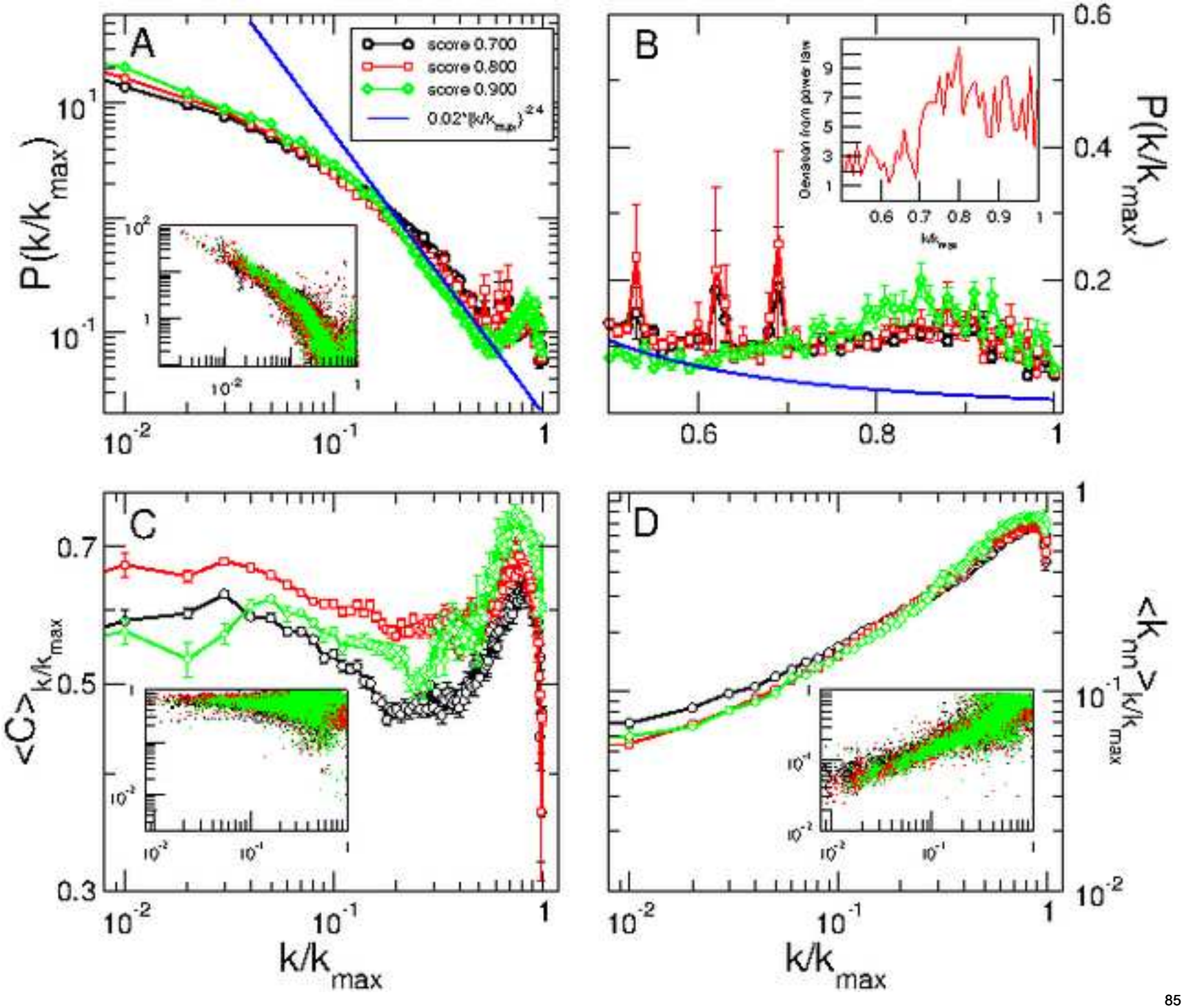


Figure 2
Click here to download high resolution image

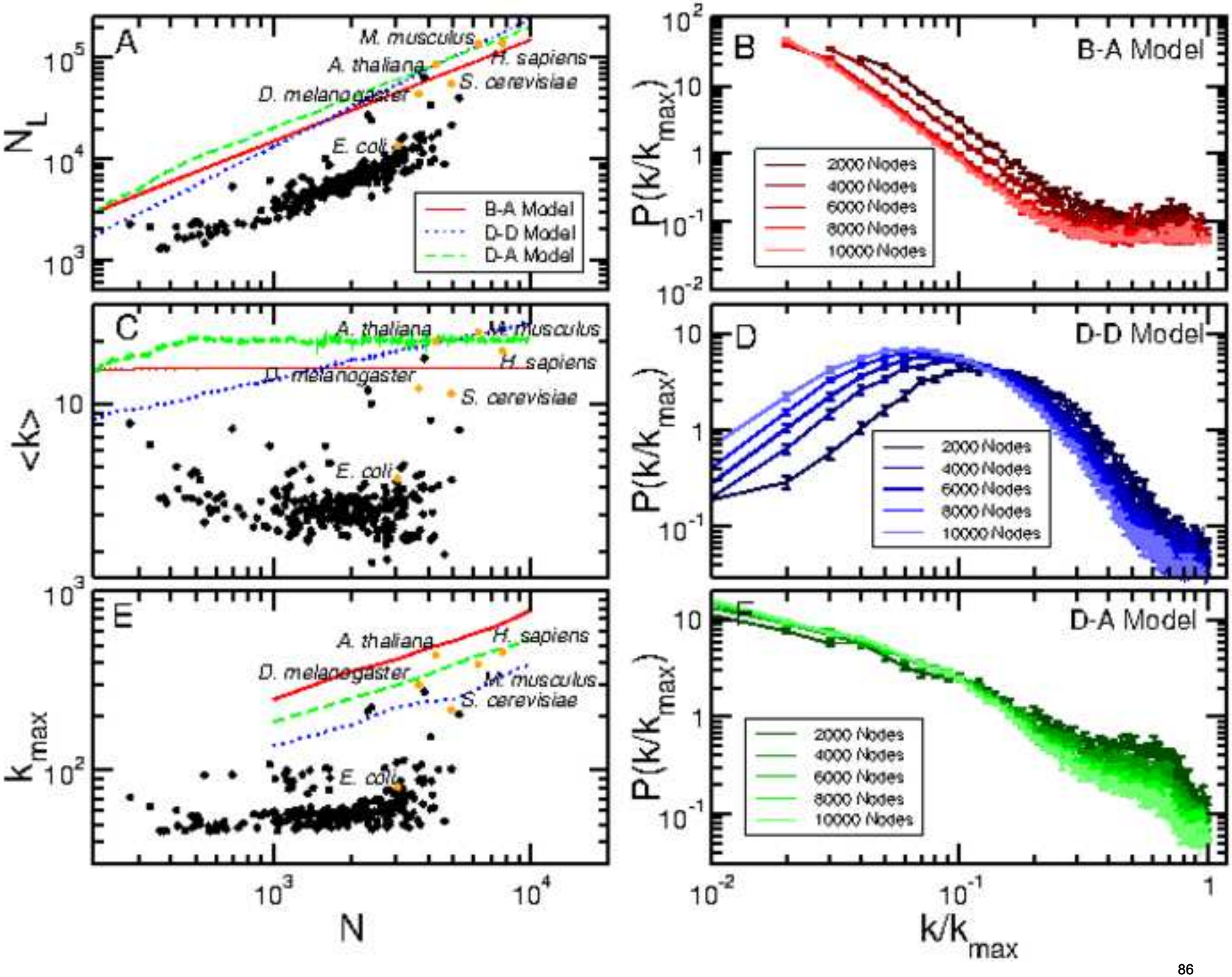


Figure 3
Click here to download high resolution image

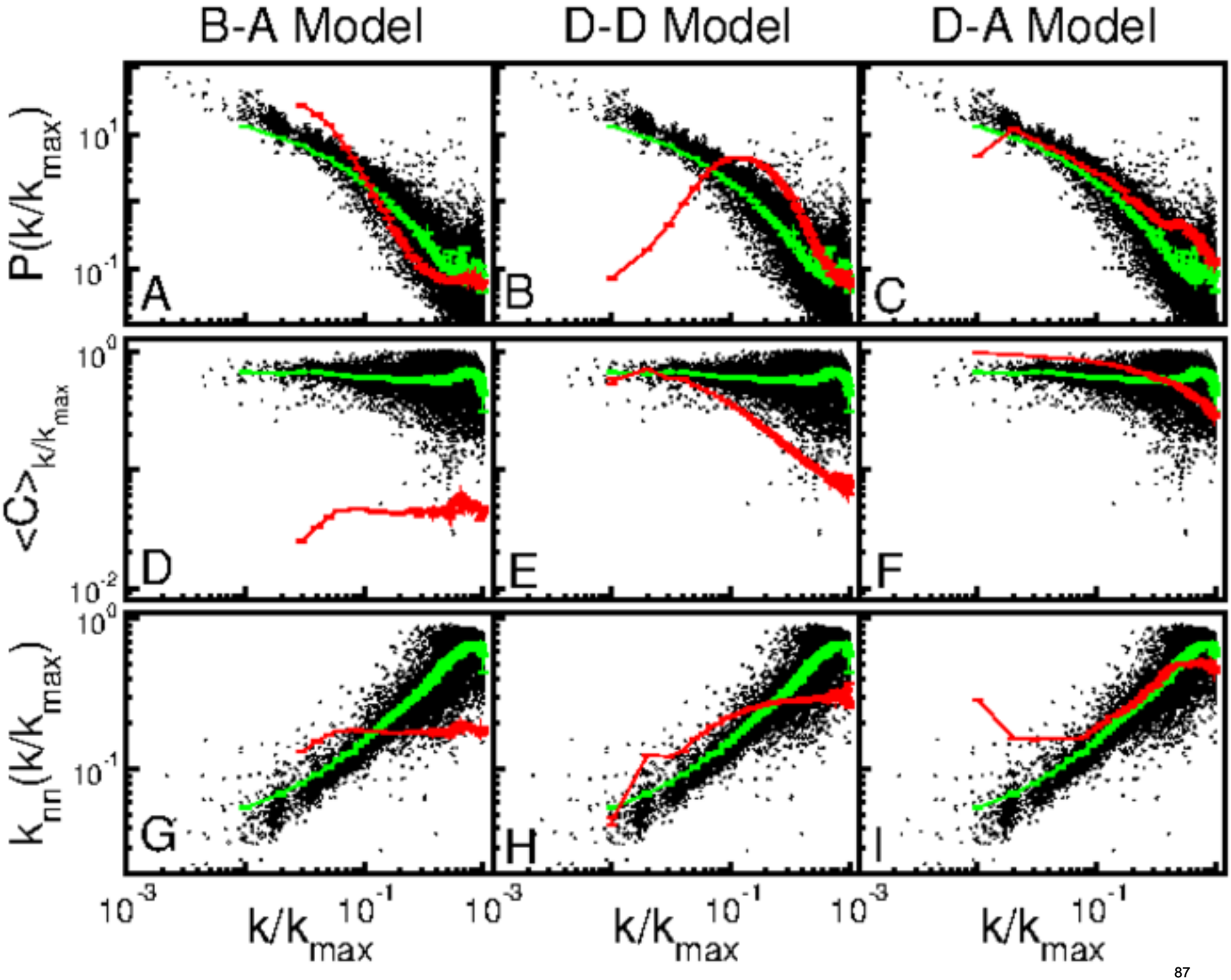
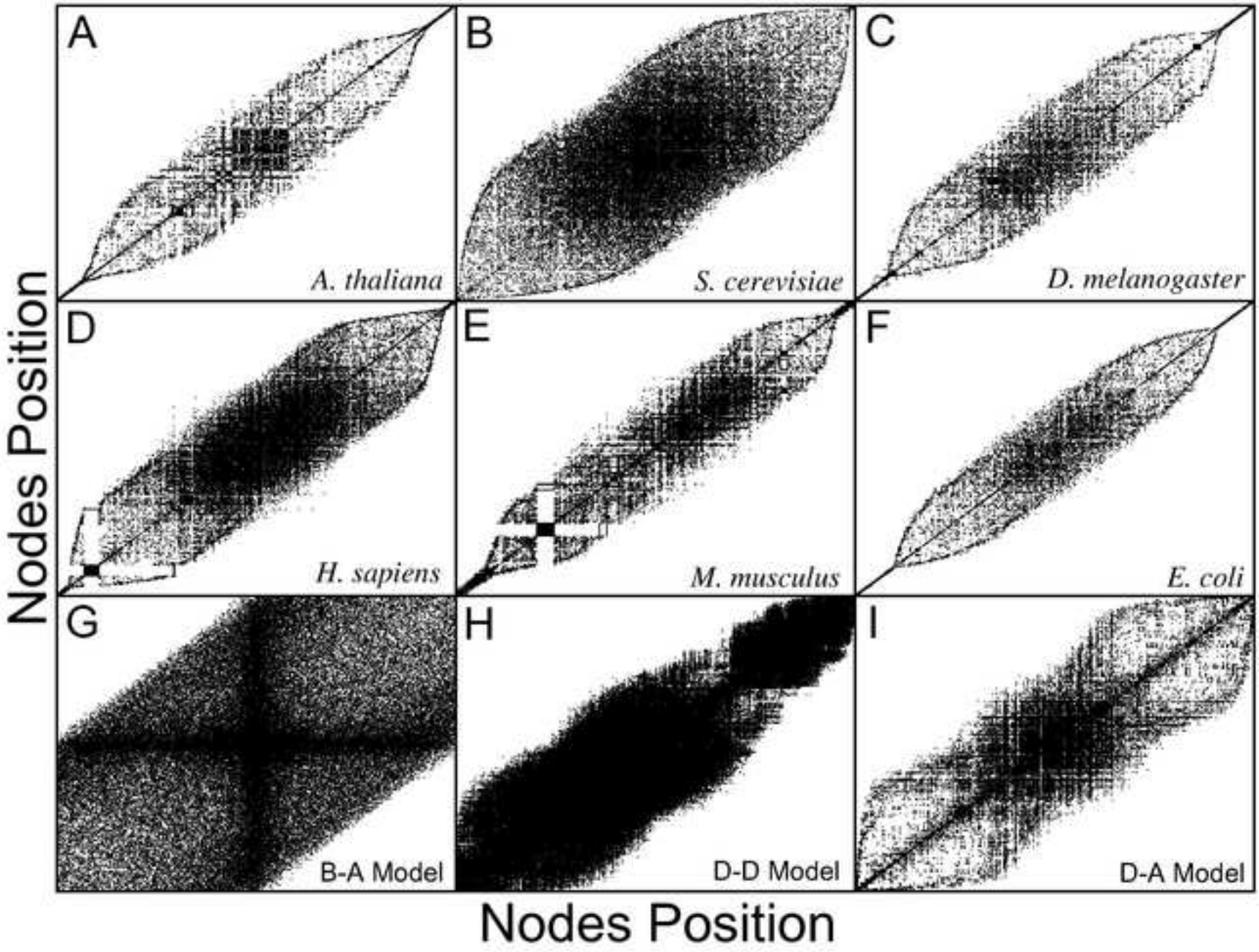


Figure 4
Click here to download high resolution image



Parte III

Discussão

Os sistemas bioquímicos presentes em organismos modernos são fruto das forças evolutivas que agiram sobre organismos ancestrais. Essas mesmas forças continuam desenhando a arquitetura dos atuais sistemas em um processo contínuo que provavelmente sempre acompanhará a vida. Entretanto, os diferentes sistemas bioquímicos atuais não surgiram simultaneamente. O arcabouço metabólico de uma célula de mamífero, por exemplo, contém rotas bioquímicas descritas em bactérias, o que denota que tais sistemas estavam presentes no ancestral comum entre mamíferos e procariotos. Em contrapartida, alguns sistemas bioquímicos são exclusivos de mamíferos, indicando que muito provavelmente tiveram o ápice de sua construção ao longo do surgimento da classe *Mammalia*. Da mesma forma, existem sistemas específicos de alguns grupos de bactérias, os quais podem ter surgido recentemente em bactérias modernas. É presumível que os processos evolutivos atuem tanto agregando novidade a sistemas ancestrais, quanto integrando sistemas que outrora exerciam funções não concatenadas.

Na presente tese, foram primeiramente estudados dois sistemas bioquímicos modelos os quais são intimamente relacionados em organismos vertebrados: apoptose e estabilidade genômica (Sengupta and Harris, 2005). Danos ao DNA, como quebra de fita simples ou quebra de fita dupla, são reconhecidos em mamíferos pelos sistemas de reparo. Uma vez identificado, o dano pode ser corrigido pelos sistemas de reparo, como por exemplo, o sistema de excisão de base (BER, do inglês, *base excision repair*). Entretanto, quando o dano é extenso demais para ser reparado, o sistema de apoptose é ativado, eliminado assim a célula com o material genético danificado (Rios and Puhalla, 2011). Diversas rotas bioquímicas são relacionadas com o reparo de DNA e com a

estabilização do genoma. Na presente tese, elas são conjuntamente denominadas como sistema de estabilidade genômica. Castro e colaboradores evidenciaram a intrincada relação entre os mecanismos de apoptose e as diferentes rotas envolvidas com a estabilização do genoma humano (estabilidade cromossômica, reparo recombinacional, reparo de *mismach*, BER e NER), em uma rede de interações proteína-proteína, envolvendo ao todo 180 genes (Castro et al., 2007).

Apesar da estreita relação funcional entre apoptose e estabilidade genômica observada em humanos, tais sistemas apresentam uma história evolutiva bastante distinta. Os sistemas de reparo de DNA têm sido extensamente descritos em procariotos, sugerindo que surgiram muito cedo na escala evolutiva, provavelmente associados aos primeiros organismos celulares (Helling, 1968; Setlow and Carrier, 1964; Wildenberg and Meselson, 1975; Willetts and Clark, 1969). Em contrapartida, apoptose compreende um sofisticado sistema bioquímico o qual atua basicamente na eliminação de células que perderam sua funcionalidade, seja por excesso de mutações, seja pelo fato de não serem mais necessárias ao organismo. Embora mecanismos de morte celular programada sejam descritos tanto em procariotos como em eucariotos unicelulares, sistemas elaborados de morte celular, como é o caso da apoptose, são associados ao surgimento de organismos pluricelulares (Ameisen, 2002).

Os resultados apresentados no capítulo I da presente tese corroboram o observado na literatura, mostrando que o sistema de estabilização do genoma estava presente na origem dos eucariotos. Em relação ao sistema de apoptose, alguns poucos componentes possuem ortólogos presentes na base da árvore eucariótica. Entretanto, boa parte dessas proteínas apoptóticas identificadas como tendo surgido no início da evolução dos eucariotos apresentam outras funções biológicas além de participarem da maquinaria de apoptose. Este é o caso de citocromo c, que apresenta reconhecida função

na cadeia transportadora de elétrons. É possível que tais componentes exercessem suas diferentes funções em sistemas bioquímicos não necessariamente relacionados à apoptose, tendo sido exaptados para essa nova função. De acordo como nossos resultados, a maioria dos componentes do sistema de apoptose encontrado em humanos foi sequencialmente recrutada ao longo da evolução dos eucariotos. No cenário proposto, dois eventos evolutivos merecem ser ressaltados: o surgimento dos metazoários, onde encontramos a origem de vários ortólogos de proteínas da via intrínseca, e o surgimento dos vertebrados, onde encontramos o surgimento de vários ortólogos de proteínas da via extrínseca.

De acordo com o panorama exposto aqui, o surgimento da apoptose parece ter sido relacionado a um controle da ontogenia de organismos pluricelulares. Uma característica do desenvolvimento dos metazoários é a produção de células em excesso durante o desenvolvimento. Células essas que, quando não são mais necessárias, devem ser eliminadas, principalmente durante os últimos estágios do desenvolvimento. De fato, este é um importante papel do mecanismo de apoptose que auxilia o organismo em desenvolvimento a alcançar um número de células compatível com o funcionamento adequado dos órgãos e tecidos (Meier et al., 2000). Apesar da possibilidade da ligação entre os sistemas de apoptose e estabilidade genômica remeter a este período evolutivo, os resultados aqui discutidos não são suficientes para inferirmos um funcionamento em conjunto dos dois sistemas já na origem dos metazoários. É possível que tanto a apoptose - ainda basicamente composta pela via intrínseca - quanto os mecanismos de estabilidade genômica, funcionassem de forma independente neste período e que a intrincada relação entre ambos tenham coevoluído subsequentemente durante a evolução.

Organismos multicelulares simples, compostos quase que totalmente por células somáticas pós-mitóticas, raramente desenvolvem tumores. Isso pode ser exemplificado por organismos como *Drosophila melanogaster* e *Caenorhabditis elegans*, descritos como organismos que não desenvolvem cânceres. Em contrapartida, organismos multicelulares complexos, compostos por células pós-mitóticas e tecidos renováveis, são propensos a desenvolverem câncer (Campisi, 2008). Graças a isso, os mecanismos supressores de tumor cresceram de importância após o advento dos vertebrados. Este período evolutivo remete ao incremento da via extrínseca de apoptose, caracterizado pelo surgimento de proteínas ortólogas aos membros da família TNF, bem como da proteína p53. A via extrínseca de apoptose pode ter representado a adaptação de um sistema, o qual originalmente evoluiu no controle da ontogenia, a uma nova função: o controle antitumoral. É possível que neste período evolutivo os sistemas de apoptose e estabilidade genômica tenham alcançado uma topologia de rede semelhante à observada hoje em humanos.

Em suma, a evolução dos sistemas de apoptose e estabilidade genômica pode ser representada da seguinte forma: (*i*) o surgimento do sistema de estabilidade genômica na origem dos eucariotos, ou até mesmo antes disso; (*ii*) o surgimento da via intrínseca de apoptose na origem dos metazoários, provavelmente relacionada ao controle ontogenético; e (*iii*) o incremento do sistema de apoptose com o surgimento da via extrínseca na origem dos vertebrados, provavelmente relacionado ao controle antitumoral. É possível que esta dinâmica, onde novos nós são agregados a sistemas antigos adaptando-os, ou mesmo criando novos sistemas, seja comum na evolução dos sistemas bioquímicos.

Uma característica importante discutida no capítulo I da presente tese refere-se à grande conservação da rede de estabilidade genômica. Além da grande maioria das suas

proteínas apresentarem ortólogos que remetem à origem dos eucariotos, quase nenhum episódio de duplicação gênica parece ter sido fixado ao longo da história destas famílias de genes. É possível que o sistema seja tão pouco tolerante a modificações, que qualquer alteração pode diminuir drasticamente a adaptabilidade do sistema como um todo, sendo fortemente constrangida pelas forças evolutivas. Essa ideia é reforçada pela essencialidade dos componentes do sistema de estabilidade genômica, demonstrada em *Saccharomyces cerevisiae, Mus musculus* e *Homo sapiens*. Os resultados observados dão conta que esta rede provavelmente alcançou sua topologia muito cedo na evolução dos eucariotos. Isso pode explicar a intrincada relação entre os seus componentes, onde a maioria dos nós são fortemente interconectados, caracterizando um módulo funcional (Barabasi and Oltvai, 2004; Castro et al., 2007).

Ainda de acordo com o nosso modelo, o sistema de apoptose compreende a região mais recente da rede. E, ao contrário do observado no sistema de estabilidade genômica, a apoptose congrega a plasticidade da rede. A partir da análise da distribuição dos grupos de ortólogos que compõe o sistema de apoptose, podemos concluir que a porção de maior plasticidade corresponde às proteínas da via extrínseca de apoptose, justamente o mecanismo que evolui mais recentemente. De fato, há uma correlação entre plasticidade e ancestralidade em toda a rede de apoptose e estabilidade genômica, onde a porção mais ancestral é a menos plástica e a mais essencial. Ao contrário, a região mais recente é a mais plástica e a menos essencial, ao menos em nível celular. É possível que o sistema de estabilidade genômica tenha alcançado um platô evolutivo, ao passo que o sistema de apoptose está em meio a sua evolução. Qualquer alteração que ocorra no primeiro, portanto, terá maior probabilidade de ser negativamente selecionada. Já o segundo, é mais tolerante a modificações.

Os resultados até aqui discutidos indicam um cenário evolutivo onde ancestralidade e plasticidade são inversamente proporcionais. Sugerem também uma relação inversa entre plasticidade evolutiva e essencialidade. Do ponto de vista topológico, a porção mais ancestral e menos tolerante a alterações apresentou-se intrincadamente conectada (denominada de módulo ρ). Entretanto, essas observações foram feitas em uma rede composta por 180 genes de uma única espécie. É possível que estas correlações possam ser encontradas de forma geral em genomas de diferentes organismos, mas esta seria uma conclusão precipitada se baseada exclusivamente nos resultados apresentados até aqui. A fim de verificar o quanto este cenário é coerente e robusto ao ponto de permitir a proposta de um mecanismo evolutivo, partimos para uma investigação global dessas correlações.

O capítulo II da presente tese apresenta uma extensa investigação acerca da plasticidade evolutiva envolvendo a distribuição dos componentes de grupos de ortólogos ao longo dos eucariotos. Foram investigados todos os KOGs presentes no repositório STRING, versão 8.2, totalizando 4850 KOGs. Ao todo, a análise envolveu 481421 proteínas em 55 eucariotos. Quanto à distribuição de abundância e diversidade, de maneira geral, o mesmo padrão observado na rede de apoptose e estabilidade genômica foi encontrado ao analisarmos todos os KOG presentes no repositório STRING. Isso é um indicativo de que, ao menos do ponto de vista da plasticidade evolutiva, a rede de apoptose e estabilidade genômica pode ser considerada representativa.

A observação de que existem grupos de ortólogos que são distribuídos ao longo da árvore dos eucariotos e que não apresentam registro da fixação de genes duplicados, nos levou a propor um índice baseado nesses dois fatores. O índice de plasticidade evolutiva - *EPI* (do inglês, *Evolutionary Plasticity Index*), considera como conservados

grupos de ortólogos que estejam amplamente distribuídos entre os eucariotos e que apresentem poucas duplicações gênicas em sua história evolutiva. Este índice, como o nome sugere, é baseado na plasticidade evolutiva de um grupo de ortólogos. Em outras palavras, procura diferenciar famílias de genes que sofreram mais alterações ao longo de sua história evolutiva, de famílias que sofreram menos alterações durante a evolução. É razoável supor que os genes pertencentes a grupos de ortólogos com baixa plasticidade evolutiva participem de sistemas bioquímicos igualmente pouco plásticos e vice versa. Dessa forma, pode haver uma correlação entre plasticidade evolutiva e plasticidade genética, assim como observado na rede de apoptose e estabilidade genômica.

Em uma análise global de essencialidade, nós verificamos que genes considerados letais, tanto em *Mus musculus*, quanto em *Saccharomyces cerevisiae*, têm maior probabilidade de pertencerem a KOGs de menor plasticidade. Em contrapartida, genes os quais são considerados não letais, apresentam maior probabilidade de pertencerem a grupos de ortólogos de maior plasticidade. Estes resultados seguem o mesmo padrão das observações feitas na rede de apoptose e estabilidade genômica em relação à plasticidade e à letalidade. Tal observação reforça a ideia de que genes que fazem parte de grupos de ortólogos que sofreram poucas alterações ao longo de sua evolução têm maior probabilidade de estarem envolvidos com sistemas essenciais, pouco tolerantes a mudanças, em organismos modernos. Ou seja, podemos inferir aspectos funcionais de um gene a partir da história evolutiva do grupo de ortólogos do qual ele faz parte.

De modo geral, a relação inversa entre plasticidade e essencialidade observada na rede de apoptose e estabilidade genômica no capítulo I foi encontrada globalmente no capítulo II. Algumas famílias de genes são menos tolerantes a alterações,

provavelmente por estarem envolvidas com sistemas que estão próximos do seu platô evolutivo. Alguns trabalhos discutem que a taxa evolutiva de um gene é inversamente proporcional a sua idade (Alba and Castresana, 2005). Essa hipótese sugere que, à medida que um gene estabiliza-se em um sistema, aumenta a constrição e a taxa evolutiva cai. Dessa forma, os sistemas mais antigos estariam menos propensos a alterações.

Este parece ser o caso do sistema de estabilidade genômica, discutido no capítulo I. Também parece ser o caso do ribossomo, do ciclo de Krebs e da cadeia transportadora de elétrons, discutidos no capítulo II. Levando em conta que a duplicação gênica é o principal evento responsável pelo surgimento de novos genes, é pouco provável que um sistema que surgiu há bilhões de anos e praticamente não apresenta registro de duplicações, como o ribossomo, possa servir como fonte de novidade genética. Dessa forma, é razoável imaginarmos que o surgimento de novidade genética ocorra em sistemas bioquímicos de maior plasticidade. Esta é uma valiosa informação no entendimento da dinâmica evolutiva dos sistemas bioquímicos, já que a identificação de onde surge a novidade genética pode auxiliar sobremaneira na compreensão das mudanças sofridas pelos genomas (Zhou and Wang, 2008).

É importante ressaltar aqui dois aspectos sobre o surgimento de novidade genética. Primeiramente, é consenso que a principal fonte de matéria-prima responsável pelo surgimento de novos genes advém de eventos de duplicação gênica (Long et al., 2003). Segundo, muitos são os estudos acerca de como os genes recém-duplicados encontram novas funções, embora não haja consenso sobre o caminho a ser percorrido pelas novas cópias (Kaessmann, 2010); é bastante provável que, dependendo do caso, os genes recém-duplicados sigam caminhos diferentes de acordo com as suas características (Conant and Wolfe, 2008; Jones and Begun, 2005; Roth et al., 2007).

Entretanto, poucos trabalhos discutem quais características genéticas aumentariam a probabilidade de fixação de um episódio de duplicação. Ademais, muitas vezes tais trabalhos apresentam resultados contraditórios. Em *Saccharomyces cerevisiae*, o aumento na probabilidade de duplicação tem sido relacionado com genes menos importantes (He and Zhang, 2006a). Contrariamente, essa relação não foi encontrada em *Mus musculus* (Liao and Zhang, 2007). De fato, há escassa informação acerca de onde surge a novidade genética.

Alguns trabalhos têm utilizado redes de interação na tentativa de evidenciar características que distingam genes com maior ou menor probabilidade de fixar uma duplicação. Prachumwat e Li encontraram uma correlação negativa entre duplicação e conectividade em *Saccharomyces cerevisiae* (Prachumwat and Li, 2006). Entretanto, conectividade é uma medida relativamente ambígua, já que existem ao menos dois tipos de genes muito conectados (ou *hubs*): (*i*) aqueles que fazem parte de módulos biológicos, denominados de *hubs intramodulares*, e (*ii*) aqueles que conectam módulos biológicos, chamados de *hubs intermodulares*. Os hubs intramodulatres interagem com muitas proteínas simultaneamente e raramente apresentam comportamento pleiotrópico. Em contrapartida, os hubs intermodulares apresentam função pleiotrópica e conectam diferentes módulos, interagindo com diferentes parceiros em diferentes momentos e/ou em diferentes compartimentos celulares (Fraser, 2005; Han et al., 2004).

Li e colaboradores verificaram que hubs intermodulares apresentam uma maior taxa de duplicação do que hubs intramodulares em *S. cerevisiae* (Li et al., 2006). Uma observação semelhante havia sido feita por Fraser um ano antes. Também avaliando redes de interações em *S. cerevisiae,* o autor identificou que hubs intramodulares apresentam uma menor taxa evolutiva quando comparados com hubs intermodulares. Além disso, os primeiros eram mais facilmente encontrados em COGs do que os

segundos, indicando que hubs intramodulares são mais distribuídos nas espécies avaliadas (Fraser, 2005). Ambos os autores sugerem que a novidade genética surge preferencialmente em nós que ligam módulos, ao invés de surgirem em nós que fazem parte de módulos biológicos.

Os resultados evidenciados tanto no capítulo I quanto no capítulo II, corroboram a ideia de que a novidade genética surge em hubs intermodulares. De acordo com nossos resultados, proteínas que pertencem a KOGs de baixa plasticidade tendem a se conectar a outras proteínas também pertencentes a KOGs de baixa plasticidade. Isso é verdade para as proteínas componentes dos sistemas de estabilidade genômica, ribossomo, ciclo de Krebs e cadeia transportadora de elétrons. Essa observação reforça a ideia de que tais sistemas alcançaram sua arquitetura muito cedo na evolução, sendo que são compostos largamente por proteínas que igualmente surgiram muito cedo na evolução. Em relação à topologia destes sistemas, todos apresentarem características modulares, corroborando os resultados apresentados por Li e Fraser (Fraser, 2005; Li et al., 2006).

No capítulo III, efetuamos uma ampla investigação acerca das propriedades topológicas das redes biológicas, envolvendo 268 organismos. Foram investigadas as distribuições de conectividade, coeficiente de clusterização e conectividade média dos vizinhos. A partir dos dados de conectividade normalizados pela conectividade máxima encontrada em cada espécie, verificamos que a distribuição do grau não corresponde a uma lei de potência. Na verdade, apresenta uma dinâmica ímpar, observada de modo geral nas espécies analisadas, que indica um aumento na probabilidade de encontrarmos um gene altamente conectado. Esses resultados indicam que proteínas com alta conectividade tendem a surgir nas redes proteicas. Esta é uma dinâmica diferente da observada em trabalhos anteriores (Barabasi and Oltvai, 2004). Tal comportamento de

conectividade foi observado pela primeira vez, devido principalmente à normalização dos dados pela conectividade máxima de cada espécie. Isso permitiu uma correta comparação dos dados topológicos de redes proteicas de diferentes espécies. A partir dessas observações e, em conjunto com o observado nos capítulos I e II, propusemos um modelo computacional de crescimento de rede levando em conta as condições discutidas a seguir.

A partir de uma rede semente, novos nós foram adicionados por duplicação, ou seja, cada novo nó herdou as ligações de seu nó parental. Após a duplicação, algumas ligações eram perdidas no nó duplicado ou no seu parental. O objetivo desta etapa da simulação foi de mimetizar os principais mecanismos conhecidos por atuarem no surgimento de novidade genética, ou seja, o processo de duplicação gênica seguido da mutação de ao menos uma das cópias. Essa ideia já havia sido proposta por Vázquez e colaboradores (Vázquez, 2003; Vázquez et al., 2003). Entretanto, no modelo aqui discutido, os nós não foram aleatoriamente escolhidos para duplicarem. Quanto maior a conectividade e quanto menor o coeficiente de clusterização, maior a probabilidade de duplicação. Esta probabilidade procurou reproduzir a condição dos genes considerados hub intermodulares (Fraser, 2005; Li et al., 2006). 90% dos nós foram incluídos na rede a partir do mecanismo de duplicação. Os 10% restantes foram adicionados à rede seguindo um padrão de surgimento *de novo*. O objetivo foi mimetizar outros mecanismos responsáveis pelo surgimento de novidade genética, como por exemplo, a transferência horizontal de genes.

A rede obtida segundo as regras aqui discutidas apresentou diversas características similares às redes biológicas, como as distribuições de conectividade, coeficiente de clusterização e conectividade média dos vizinhos. Além da distribuição de conectividade, o modelo proposto foi capaz de descrever simultaneamente

características topológicas das redes biológicas, como a clusterização e a assortatividade. Até onde vai o nosso conhecimento, nenhum outro modelo de crescimento de redes conseguiu mimetizar de forma tão eficiente tais características das redes biológicas. Isso indica que o modelo apresentado provavelmente está mais próximo da descrição dos mecanismos de evolução reais que atuam sobre os genomas, do que outros modelos anteriormente propostos. Certamente este modelo simplificado, com um número pequeno de variáveis, não congrega todos os aspectos que regem o crescimento dos genomas. Tampouco se propõe a explicar todas nuances do processo evolutivo. Entretanto, nossos resultados indicam fortemente que a duplicação de genes, os quais codificam proteínas altamente conectadas, e que fazem a ligação entre diferentes módulos biológicos, representa um importante mecanismo no surgimento de novidade genética. É possível ainda que este seja o mecanismo majoritário de ampliação do genoma.

Essa ideia vem ao encontro de resultados obtidos anteriormente (Fraser, 2005; Li et al., 2006). Os hubs intermodulares apresentam a pleiotropia como característica funcional. O efeito pleiotrópico de um gene, em geral, não está ligado a múltiplos domínios, mas sim à utilização de uma única função molecular em diversos processos biológicos. Do ponto de vista topológico, genes pleiotrópicos apresentam-se altamente conectados a módulos funcionais diferentes (He and Zhang, 2006b). Muitas vezes, genes pleiotrópicos apresentam uma baixa taxa evolutiva (Davis and Petrov, 2004; Makino and Gojobori, 2006). Isso pode ocorrer devido a um conflito adaptativo de um gene que exerce mais de uma função. Uma possível especialização em uma função particular pode acarretar em um decréscimo funcional nas outras funções da proteína (Storz, 2008). Dessa forma, um evento de duplicação em um gene com essas características poderia ter uma grande probabilidade de ser fixado (Des Marais and

Rausher, 2008; Piatigorsky and Wistow, 1991). De modo geral, esses resultados corroboram o nosso modelo.

Sumarizando, primeiramente analisamos uma rede biológica (*i.e.* rede de apoptose e estabilidade genômica) que emergiu ao longo da evolução dos eucariotos. Este estudo permitiu observações importantes acerca de como sistemas que surgem em momentos diferentes, com funções não necessariamente relacionadas, podem evoluir a um trabalho em conjunto, quando isto se faz necessário ao longo da evolução. Os resultados sugeriram que sistemas antigos são pouco plásticos, altamente interconectados e essenciais. As observações também sugerem que a evolução atua principalmente unindo sistemas ancestrais.

Subsequentemente, verificamos que as observações efetuadas na rede de apoptose e estabilidade genômica em relação à plasticidade, representam um mecanismo geral, e não particular àquela rede. Observamos também que existem famílias de proteínas que praticamente não apresentam duplicações gênicas fixadas, ao passo que outras famílias de genes são altamente propensas a fixarem genes duplicados. Tais resultados sugerem que a novidade genética não surge aleatoriamente no genoma, mas que determinados grupos de genes terão maior probabilidade de servirem como fonte de surgimento de novos genes.

Finalmente, construímos um modelo de expansão do genoma onde a novidade genética surge basicamente a partir da duplicação de hubs intermodulares. A rede artificial criada a partir dessa regra, a qual segue as observações obtidas nos trabalhos anteriormente citados, mimetiza de forma bastante satisfatória a topologia das redes de interação proteína-proteína conhecidas. Os resultados aqui apresentados, quando analisados em conjunto, indicam que os processos evolutivos atuam principalmente

unindo sistemas ancestrais, agregando novidade genética à periferia de sistemas que chegaram próximos a um platô evolutivo.

Conclusões

Os estudos realizados na presente tese nos permitiram chegar às seguintes conclusões:

1- Os sistemas de estabilidade genômica e apoptose tiveram origem em momentos diferentes da evolução, sendo que o primeiro é mais antigo e menos plástico que o segundo;
2- É possível determinarmos a plasticidade evolutiva de um grupo de ortólogos a partir da análise da distribuição de suas proteínas em diferentes espécies;
3- Uma rede artificial criada a partir de um modelo matemático de duplicação gênica, onde os hubs intermodulares apresentam maior probabilidade de duplicação, descreve satisfatoriamente a topologia das redes proteicas conhecidas.

Referências Bibliográficas

Alba MM and Castresana J . 2005. Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. **Molecular Biology and Evolution** 22: 598-606.

Amaral LAN and Ottino JM . 2004. Complex networks. **The European Physical Journal B** - Condensed Matter and Complex Systems 38: 147-162.

Ameisen JC . 2002. On the origin, evolution, and nature of programmed cell death: a timeline of four billion years. **Cell Death & Differentiation** 9: 367-393.

Barabasi AL and Albert R . 1999. Emergence of scaling in random networks. **Science** 286: 509-512.

Barabasi AL and Oltvai ZN . 2004. Network biology: Understanding the cell's functional organization. **Nature Reviews Genetics** 5: 101-U15.

Campisi J . 2008. Aging and cancer cell biology, 2008. **Aging Cell** 7: 281-284.

Castro MAA, Mombach JCM, de Almeida RMC and Moreira JCF . 2007. Impaired expression of NER gene network in sporadic solid tumors. **Nucleic Acids Research** 35: 1859-1867.

Chen R and Jeong SS . 2000. Functional prediction: Identification of protein orthologs and paralogs. **Protein Science** 9: 2344-2353.

Conant GC and Wolfe KH . 2008. Turning a hobby into a job: How duplicated genes find new functions. **Nature Reviews Genetics** 9: 938-950.

Culotta VC, Klomp LWJ, Strain J, Casareno RL, Krems B and Gitlin JD . 1997. The Copper Chaperone for Superoxide Dismutase. **Journal of Biological Chemistry** 272: 23469-23472.

Dan Graur and Wen-Hsiung Li. 2000. Fundamentals of Molecular Evolution. Sunderland: Sinauer.

Davis JC and Petrov DA . 2004. Preferential Duplication of Conserved Proteins in Eukaryotic Genomes. **PLoS Biology** 2: e55.

Deng C, Cheng C-HC, Ye H, He X and Chen L . 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. **Proceedings of the National Academy of Sciences of the United States of America** 107: 21593-21598.

Des Marais DL and Rausher MD . 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. **Nature** 454: 762-765.

Fraser HB . 2005. Modularity and evolutionary constraint on proteins. **Nature Genetics** 37: 351-352.

Gelain DP, Dalmolin RJ, Belau VL, Moreira JC, Klamt F and Castro MA . 2009. A systematic review of human antioxidant genes. **Frontiers in Bioscience** 14: 4457-4463.

Glazko G and Mushegian A . 2004. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. **Genome Biology** 5: R32.

Goldenfeld N and Kadanoff LP . 1999. Simple Lessons from Complexity. **Science** 284: 87-89.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP and Vidal M . 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. **Nature** 430: 88-93.

Harrington ED, Jensen LJ and Bork P . 2008. Predicting biological networks from genomic data. **FEBS Letters** 582: 1251-1258.

He X and Zhang J . 2006a. Higher Duplicability of Less Important Genes in Yeast Genomes. **Molecular Biology and Evolution** 23: 144-151.

He X and Zhang J . 2006b. Toward a Molecular Understanding of Pleiotropy. **Genetics** 173: 1885-1891.

Helling RB . 1968. Selection of a Mutant of Escherichia coli Which Has High Mutation Rates. **Journal of Bacteriology** 96: 975-980.

Innan H and Kondrashov F . 2010. The evolution of gene duplications: classifying and distinguishing between models. **Nature Reviews Genetics** 11: 97-108.

Jones CD and Begun DJ . 2005. Parallel evolution of chimeric fusion genes. **Proceedings of the National Academy of Sciences of the United States of America** 102: 11373-11378.

Kaessmann H . 2010. Origins, evolution, and phenotypic impact of new genes. **Genome Research** 20: 1313-1326.

Kanehisa M and Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research** 28: 27-30.

Kimura M . 1991. The neutral theory of molecular evolution: A review of recent evidence. **The Japanese Journal of Genetics** 66: 367-386.

Koonin EV and Wolf YI . 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics 11: 487-498.

Koonin EV . 2005. Orthologs, paralogs, and evolutionary genomics. **Annual Review of Genetics** 39: 309-338.

Li L, Stoeckert CJ, Jr. and Roos DS . 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome Research** 13: 2178-2189.

Li L, Huang Y, Xia X and Sun Z . 2006. Preferential Duplication in the Sparse Part of Yeast Protein Interaction Network. **Molecular Biology and Evolution** 23: 2467-2473.

Liao BY and Zhang J . 2007. Mouse duplicate genes are as essential as singletons. **Trends in Genetics** 23: 378-381.

Long M and Thornton K . 2001. Gene duplication and evolution. **Science** 293: 1551.

Long M, Betran E, Thornton K and Wang W . 2003. The origin of new genes: glimpses from the young and old. **Nature Reviews Genetics** 4: 865-875.

Makino T and Gojobori T . 2006. The Evolutionary Rate of a Protein Is Influenced by Features of the Interacting Partners. **Molecular Biology and Evolution** 23: 784-789.

Mayr E . 2005. Biologia, ciência única. Companhia das Letras.

Meier P, Finch A and Evan G . 2000. Apoptosis in development. **Nature** 407: 796-801.

Nei M . 2005. Selectionism and Neutralism in Molecular Evolution. **Molecular Biology and Evolution** 22: 2318-2342.

O'Brien KP, Remm M and Sonnhammer ELL . 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. **Nucleic Acids Research** 33: D476-D480.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M . 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research** 27: 29-34.

Piatigorsky J and Wistow G . 1991. The recruitment of crystallins: new functions precede gene duplication. **Science** 252: 1078-1079.

Prachumwat A and Li WH . 2006. Protein Function, Connectivity, and Duplicability in Yeast. Molecular Biology and Evolution 23: 30-39.

Pujol A, Mosca R, Farrθs J and Aloy P . 2010. Unveiling the role of network and systems biology in drug discovery. **Trends in Pharmacological Sciences** 31: 115-123.

Rios J and Puhalla S . 2011. PARP inhibitors in breast cancer: BRCA and beyond. **Oncology** (Williston Park) 25: 1014-1025.

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D and Liberles DA . 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. **Journal of experimental zoology part B - Molecular and developmental evolution** 308: 58-73.

Rybarczyk-Filho JL, Castro MA, Dalmolin RJ, Moreira JC, Brunnet LG and de Almeida RM . 2010. Towards a genome-wide transcriptogram: the Saccharomyces cerevisiae case. **Nucleic Acids Research**.

Sengupta S and Harris CC . 2005. p53: Traffic cop at the crossroads of DNA repair and recombination. Nature Reviews Molecular **Cell Biology** 6: 44-55.

Setlow RB and Carrier WL . 1964. Disappearance of Thymine Dimers from Dna - Error-Correcting Mechanism. **Proceedings of the National Academy of Sciences of the United States of America** 51: 226-&.

Storz JF . 2008. Genome evolution: Gene duplication and the resolution of adaptive conflict. **Heredity** 102: 99-100.

Studer RA and Robinson-Rechavi M . 2009. How confident can we be that orthologs are similar, but paralogs differ? **Trends Genetics** 25: 210-216.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ and Mering Cv . 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Research** 39: D561-D568.

Tatusov RL, Koonin EV and Lipman DJ . 1997. A genomic perspective on protein families. **Science** 278: 631-637.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. **Nucleic Acids Research** 29: 22-28.

Vázquez A . 2003. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. **Physical Review E** 67: 056104.

Vázquez A, Flammini A, Maritan A and Vespignani A . 2003. Modeling of Protein Interaction Networks. **Complexus** 1: 38-44.

Wildenberg J and Meselson M . 1975. Mismatch Repair in Heteroduplex DNA. **Proceedings of the National Academy of Sciences of the United States of America** 72: 2202-2206.

Willetts NS and Clark AJ . 1969. Characteristics of Some Multiply Recombination-Deficient Strains of Escherichia coli. **Journal of Bacteriology** 100: 231-239.

Yamada T and Bork P . 2009. Evolution of biomolecular networks: lessons from metabolic and protein interactions. **Nature Reviews Molecular Cell Biology** 10: 791-803.

Zhou Q and Wang W . 2008. On the origin and evolution of new genesΓÇöa genomic and experimental perspective. **Journal of Genetics and Genomics** 35: 639-648.

Anexos

Evolutionary origins of human apoptosis and genome-stability gene networks. – Material suplementar

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

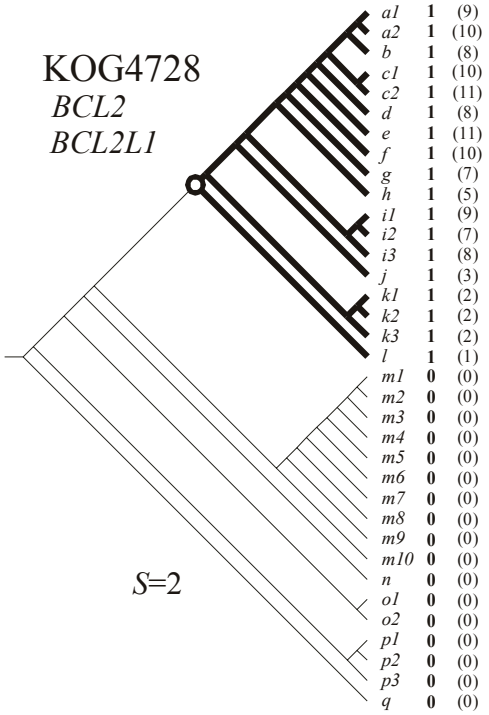
LIST OF SUPPLEMENTARY FIGURES

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

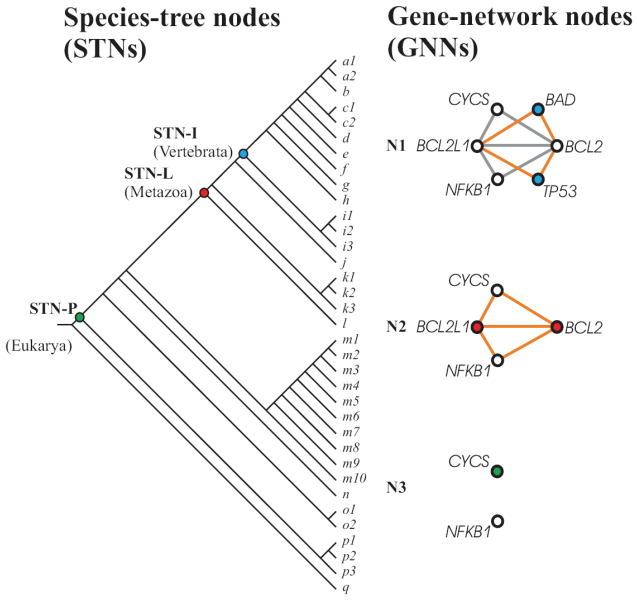
$$S = \lambda + g\gamma, \qquad (1)$$

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.

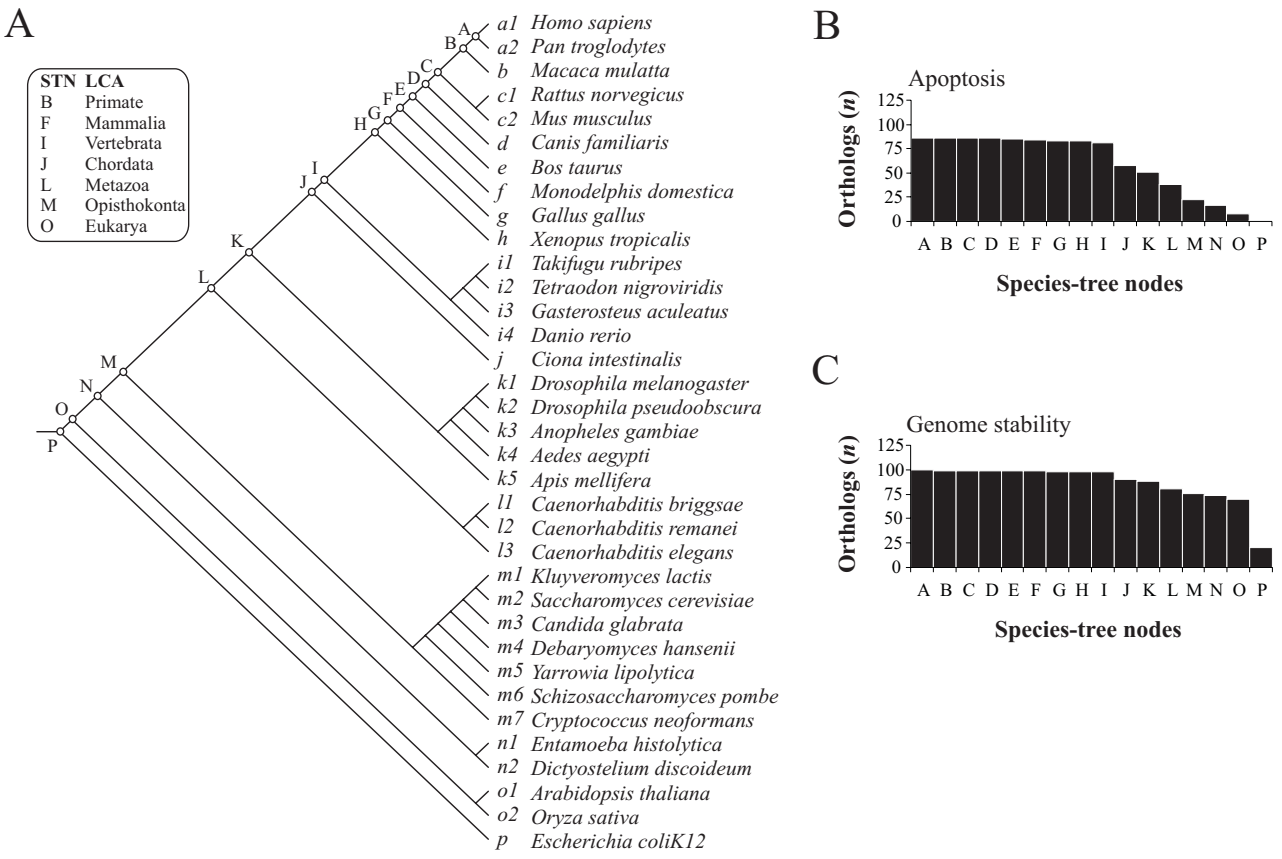
1.3. The deep root of eukaryotes

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

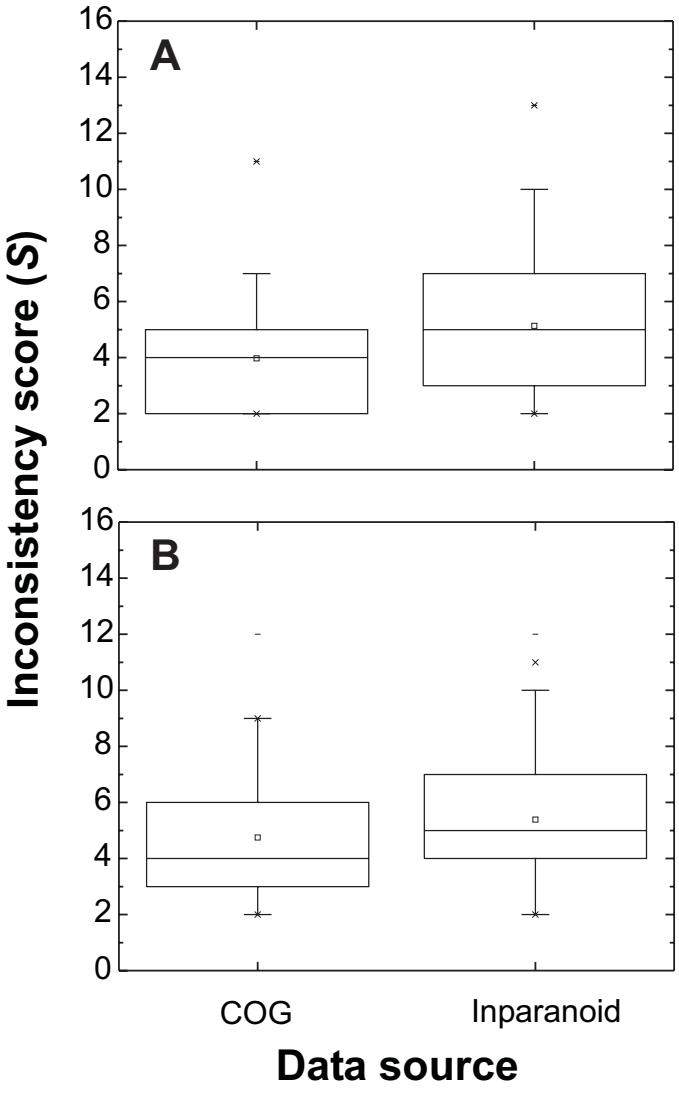
stability functions match these universally conserved COGs (**Supplementary Figure S8**, red stripes).

1.4. The deep root of metazoans

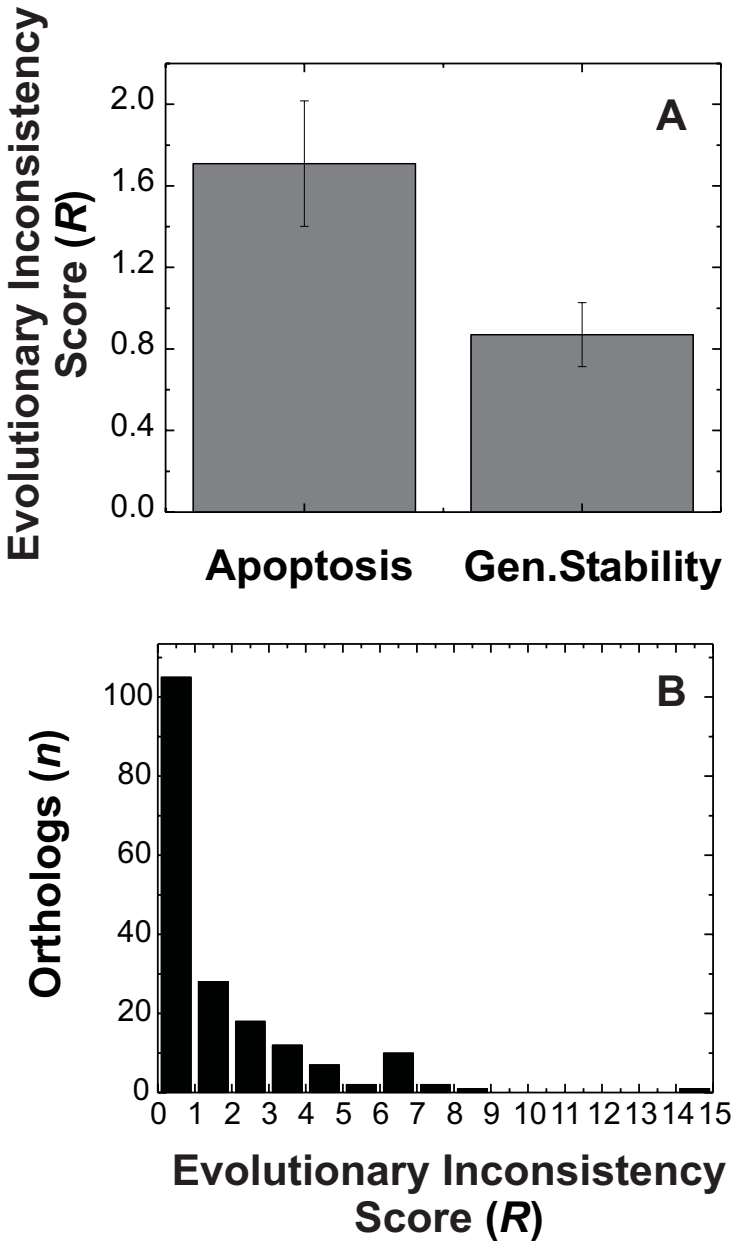
The deep root of metazoan is another controversial issue. However, the current work do not propose the phylogeny of the organisms; instead, here is considered an integration of a variety of phylogenies (Katinka et al. 2001; Pennisi 2003; Baldauf 2003; Delsuc et al. 2005; Ciccarelli et al. 2006; Letunic 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).



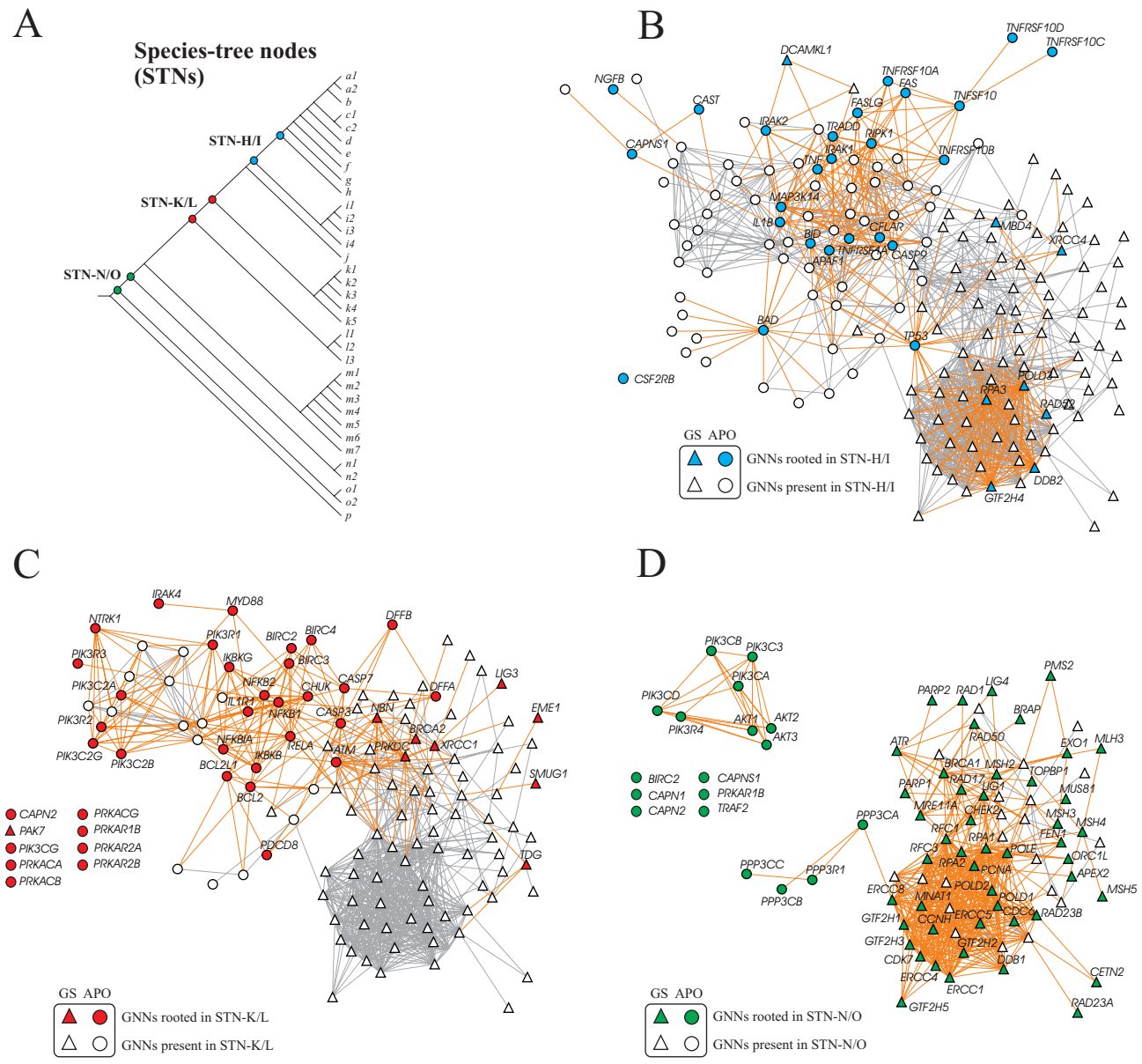
Supplementary Figure S3. Inferring ancestral state of human apoptosis and genome stability genes according to Inparanoid Database. (**A**) Eukaryote species tree topology used in the parsimony analysis. Species-Tree Nodes (STNs) and the corresponding Last Common Ancestor (LCA) are indicated. (**B**)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



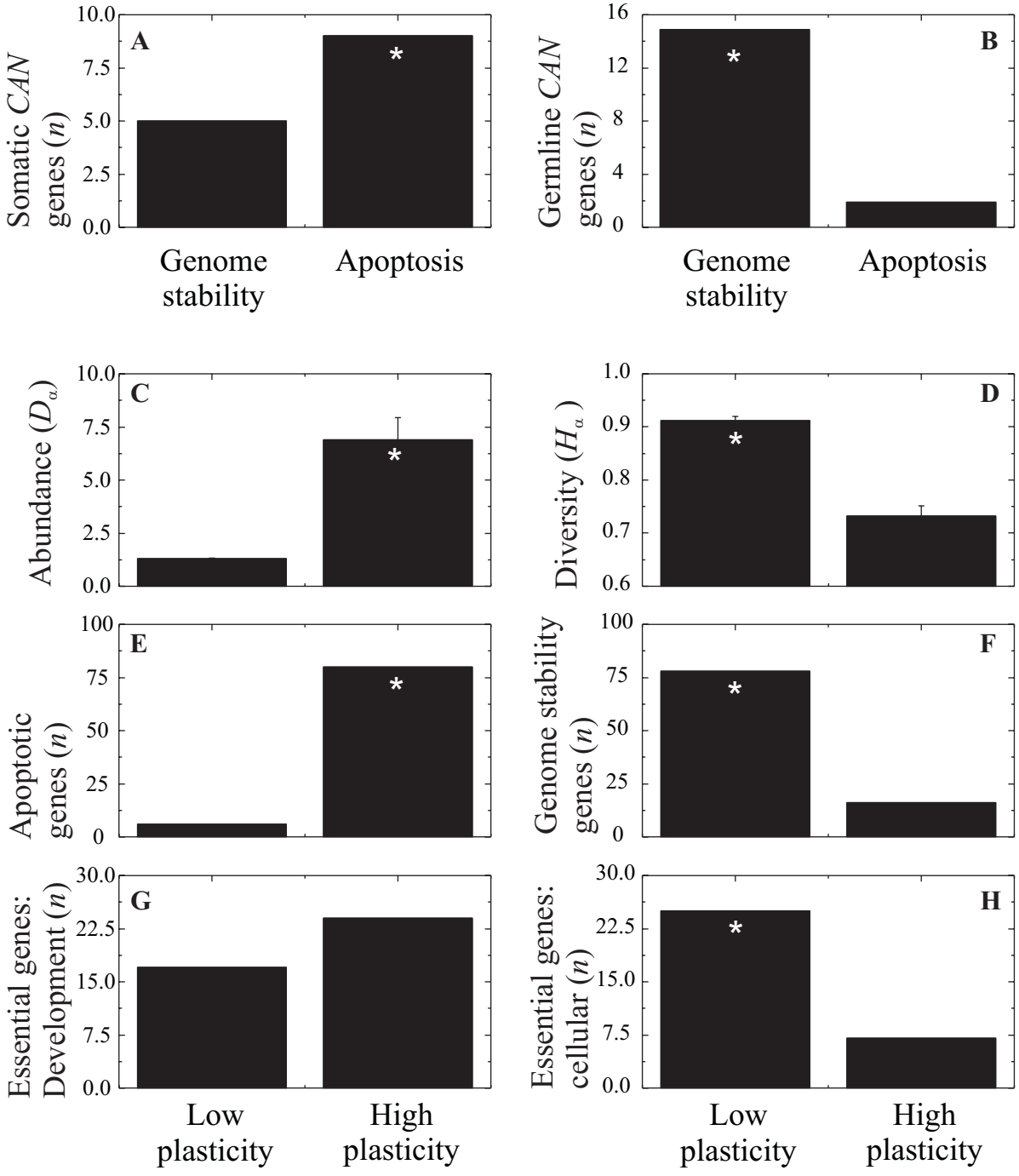
Supplementary Figure S4. Inconsistency scores *S* estimated from KOG and Inparanoid orthology analysis. Apoptosis (**A**) and genome stability gene set (**B**). For the entire gene set, S=4.39(±2.07; KOG analysis), and S=5.27(±2.48; Inparanoid analysis).



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

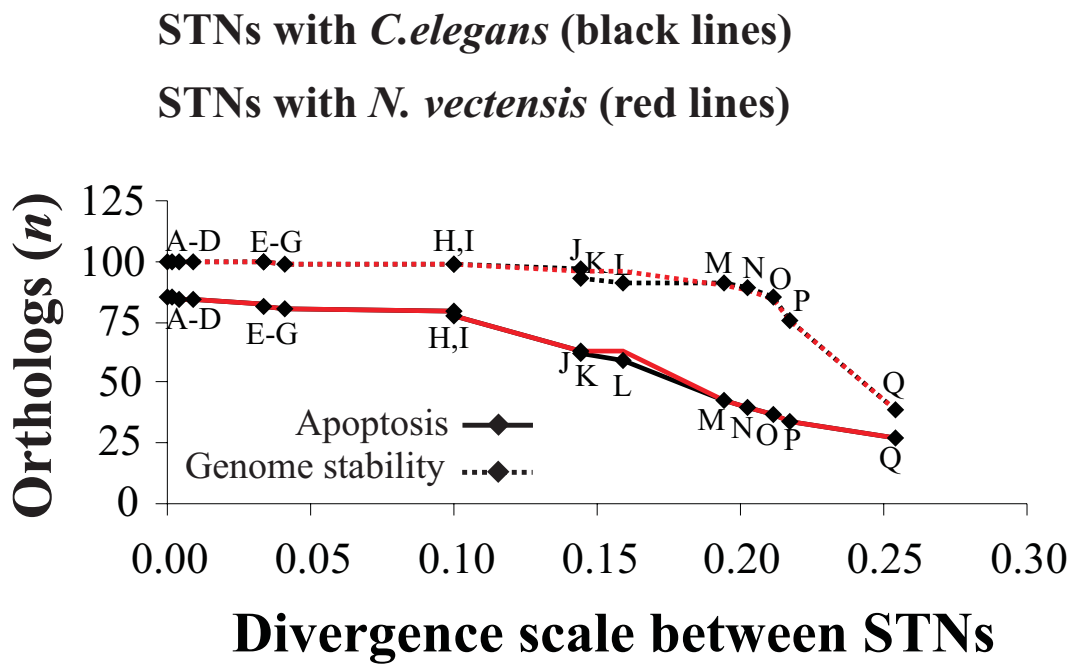
Gene Symbol	KOG-toCOG
BRAP	KOG0804
BRCA1	KOG4362
BRCA2	KOG4751
DDB1	KOG1897
EME1	NOG05923
GTF2H1	KOG2074
GTF2H5	KOG3451
MBD4	KOG4161
NBN	NOG06900
PARP1	KOG1037
PARP2	KOG1037
POLD3	NOG05166
POLD4	NOG12441
RAD1	KOG3194
RPA3	NOG10964
SHFM1	KOG4764
SMUG1	NOG04402
TOPBP1	KOG1929
XAB2	KOG2047
XRCC1	KOG3226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

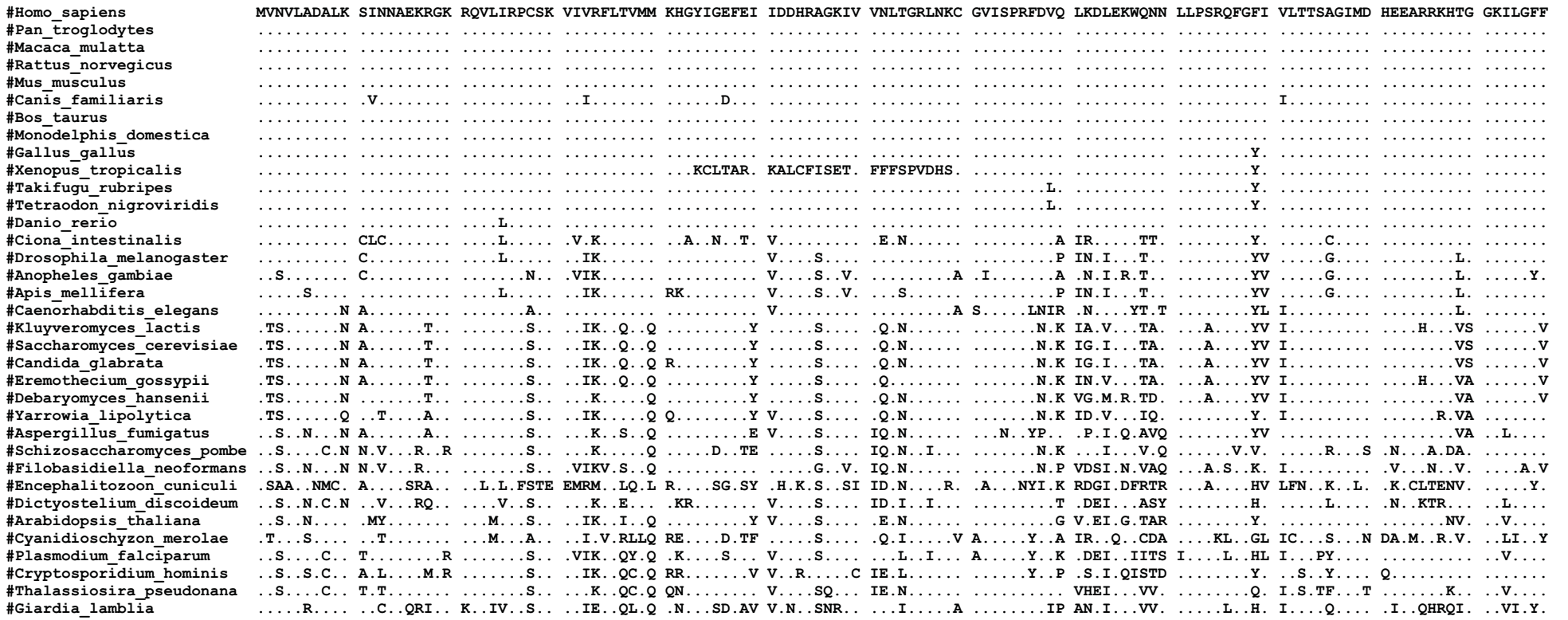
Gene Symbol	KOG-toCOG
LIG1	KOG0967 / COG1793
LIG3	KOG4437 / COG1793
LIG4	KOG0966 / COG1793
MLH1	KOG1979 / COG0323
MLH3	KOG1977 / COG0323
MNAT1	KOG3800 / COG5220
MPG	KOG4486 / COG2094
MRE11A	KOG2310 / COG0420
MSH2	KOG0219 / COG0249
MSH3	KOG0218 / COG0249
MSH4	KOG0220 / COG0249
MSH5	KOG0221 / COG0249
MSH6	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

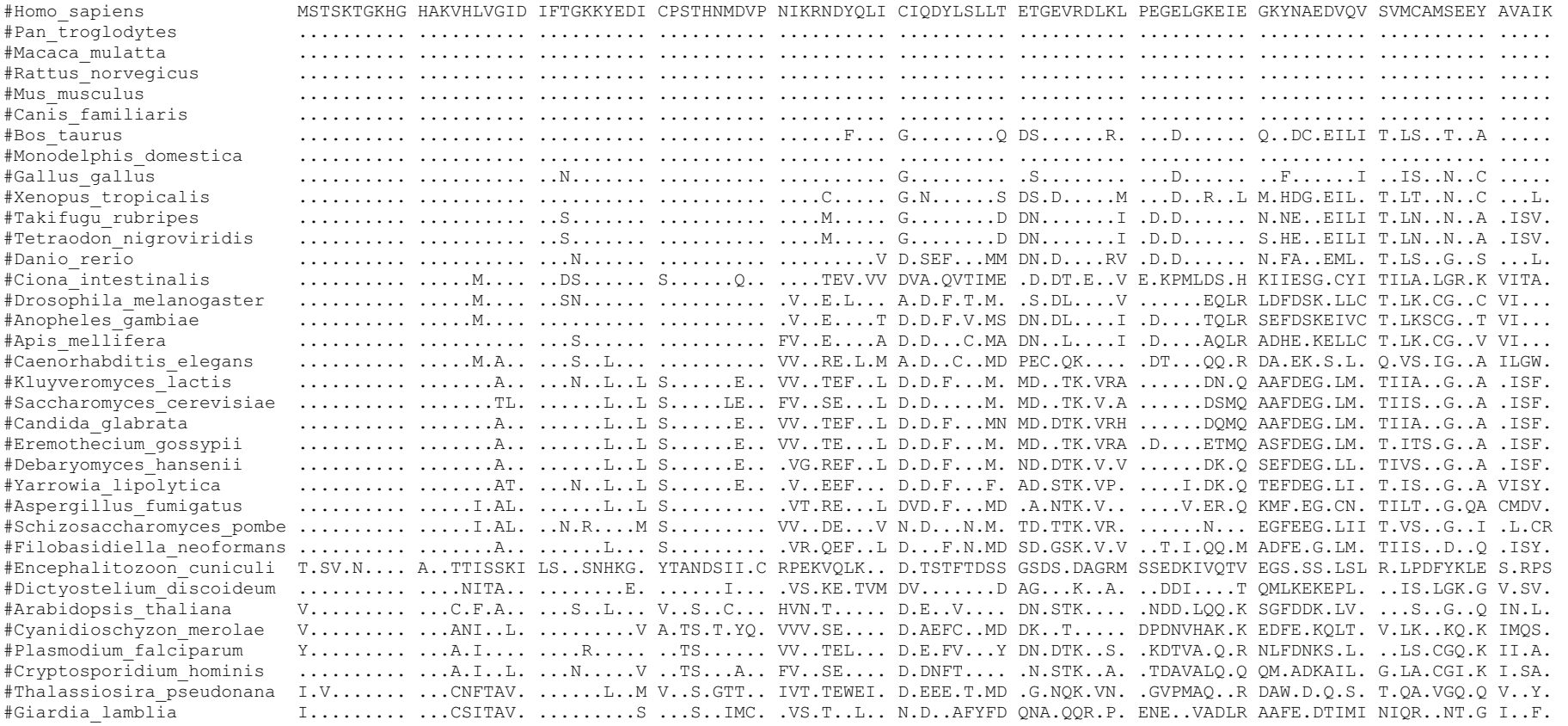
Supplementary Figure S8. KOG-to-COG correspondence. Orthology information of the Clusters of Orthologous Groups (COGs) was retrieved through the KOG-to-COG assignments in the STRING database. Red stripes highlight the eukaryotic groups that match the universally conserved prokaryotic groups in the KOG-to-COG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



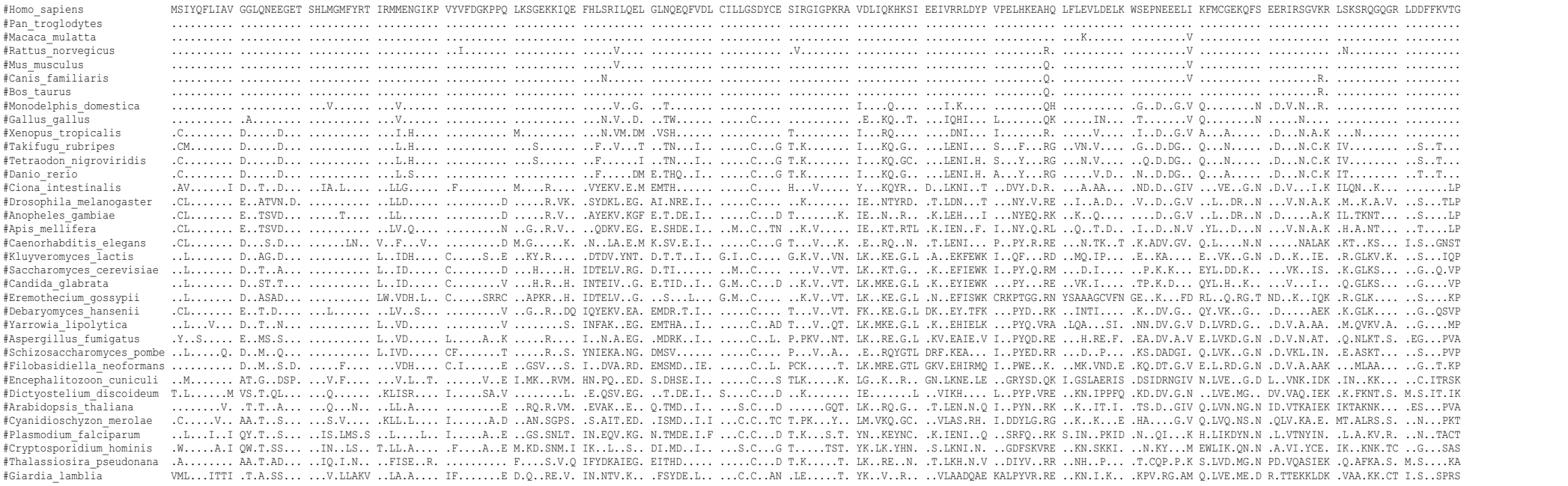
Supplementary Figure S9. Subsample of the data where *C. elegans* isn't at the root of metazoa. This figure presents a comparison between the original analysis (see **Figure 2**) and an alternative construction where *Nematostella vectensis* is manually placed in the position of *C.elegans*, according 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).



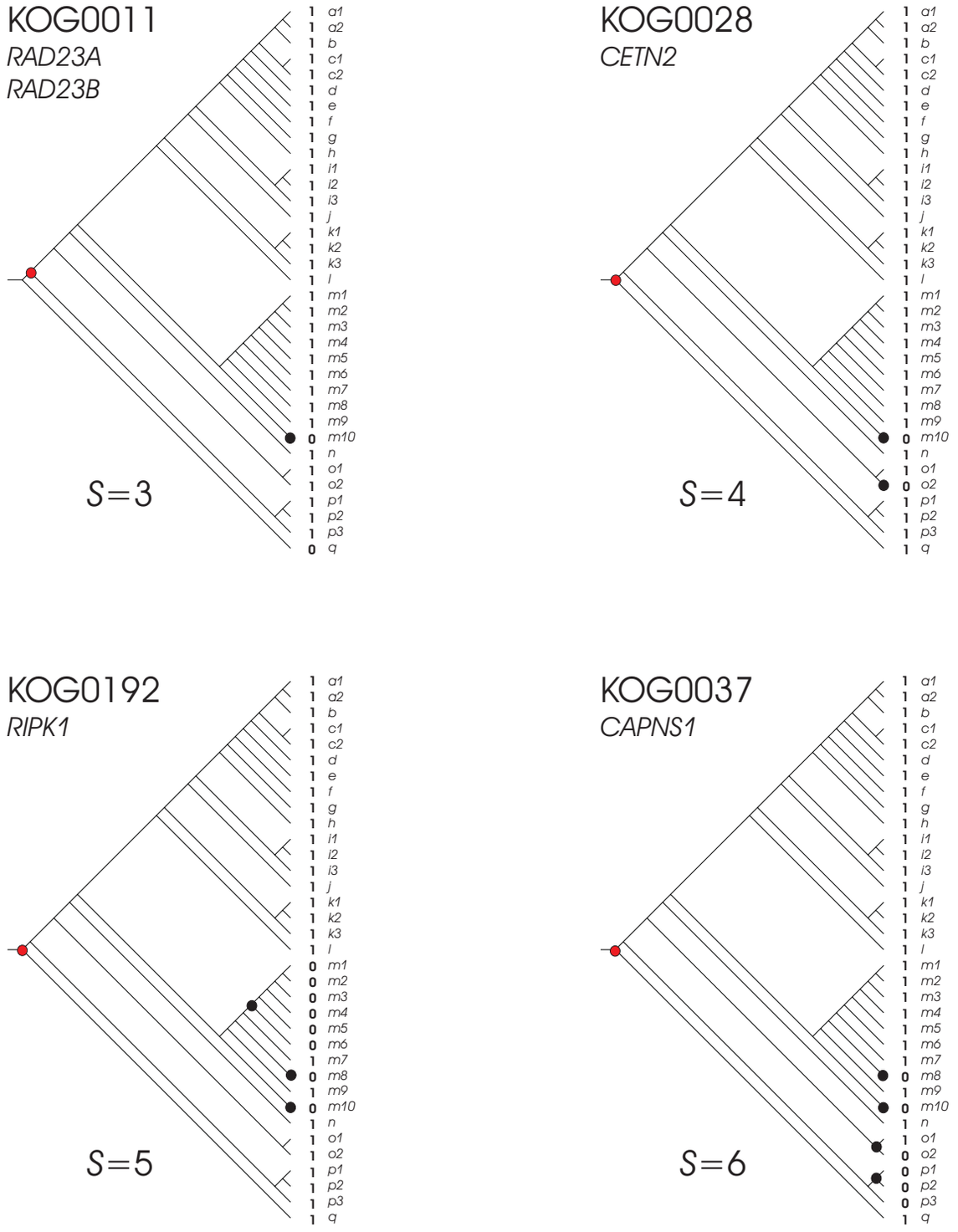
Supplementary Figure S11. Comparison of amino acid sequences of 35 eukaryotic translation initiation factor 5A proteins obtained from a three-domain orthologous groups (KOG3271/COG0231). Amino acid sequences were aligned using ClustalW. All sites containing alignment gaps and missing-information were 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).



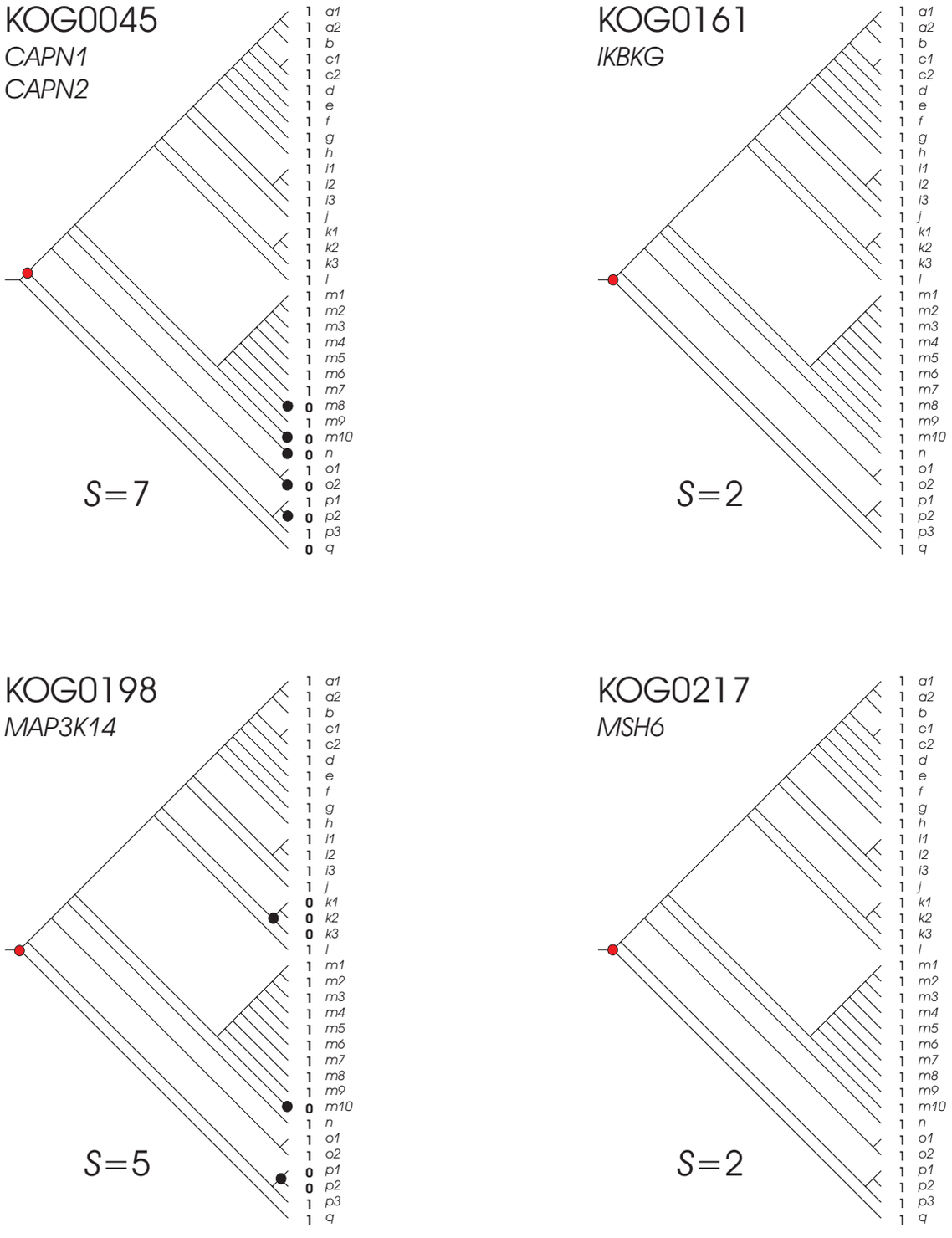
Supplementary Figure S12. Comparison of amino acid sequences of 35 eukaryotic 5-3 exonucleases (Flap Structure-specific Endonuclease 1 like proteins - FEN1) obtained from a three-domain orthologous groups (KOG2519 / COG0258). Amino acid sequences were aligned using ClustalW. All sites containing alignment 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

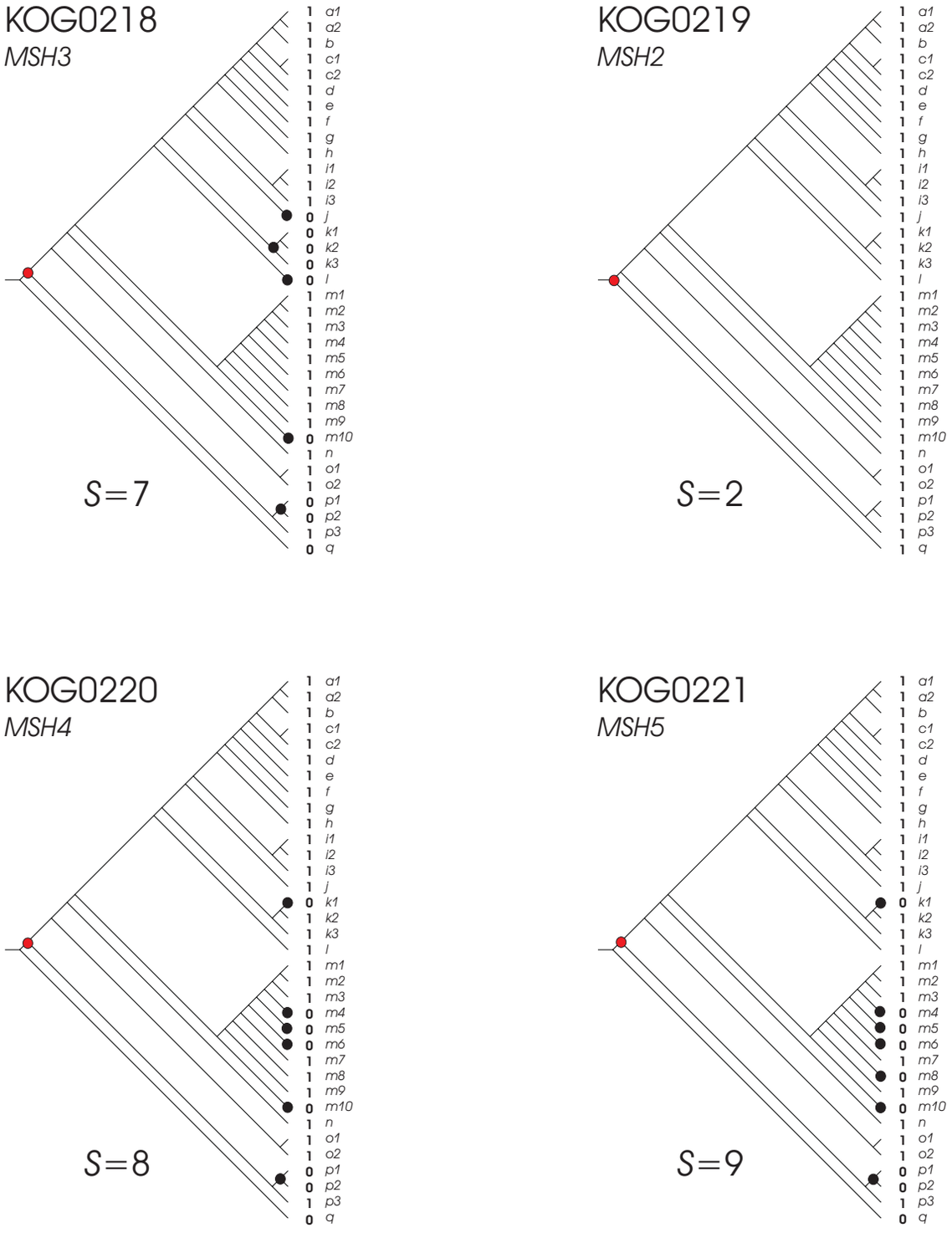
Supplementary Figure S13. Divergence matrix according to the alignments presented in **Supplementary Figures S10-S12**. The number of amino acid differences per site from analysis between sequences is shown. All results are based on the pairwise analysis of 35 sequences (average of the three proteins 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).



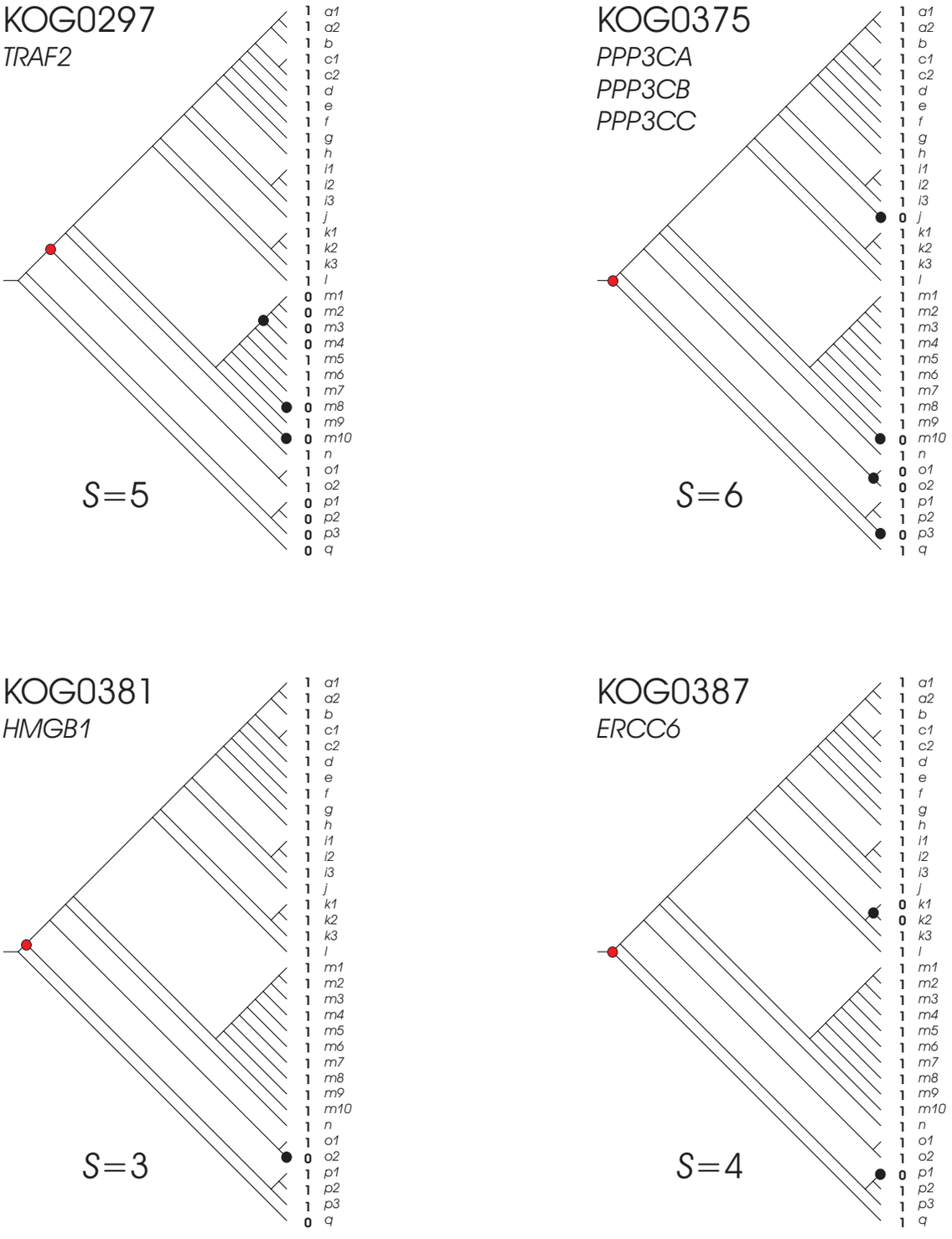
Supplementary Figure S14. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0011, KOG0028, KOG0192 and KOG0037. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (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*).



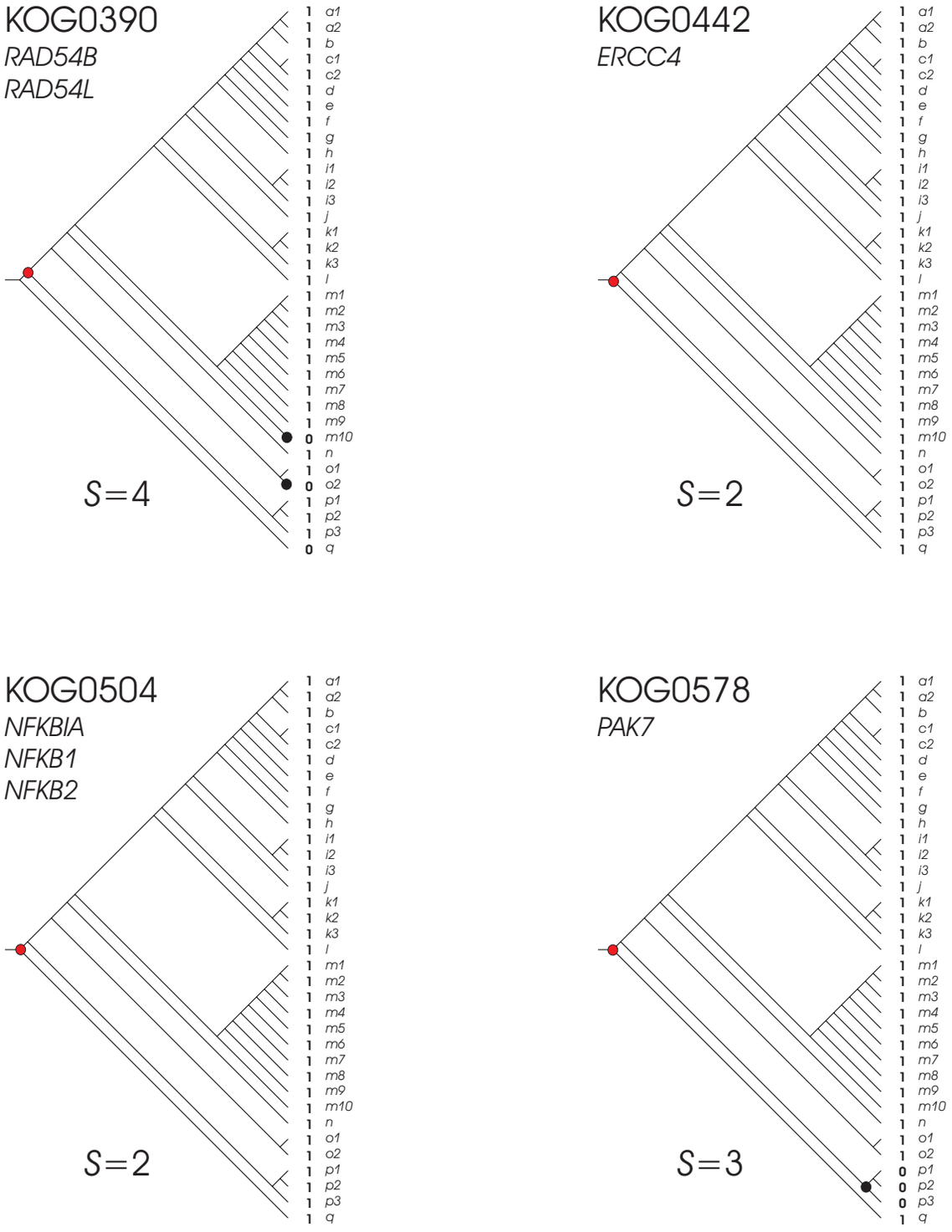
Supplementary Figure S15. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0045, KOG0161, KOG0198 and KOG0217, as in Supplementary Figure S14.



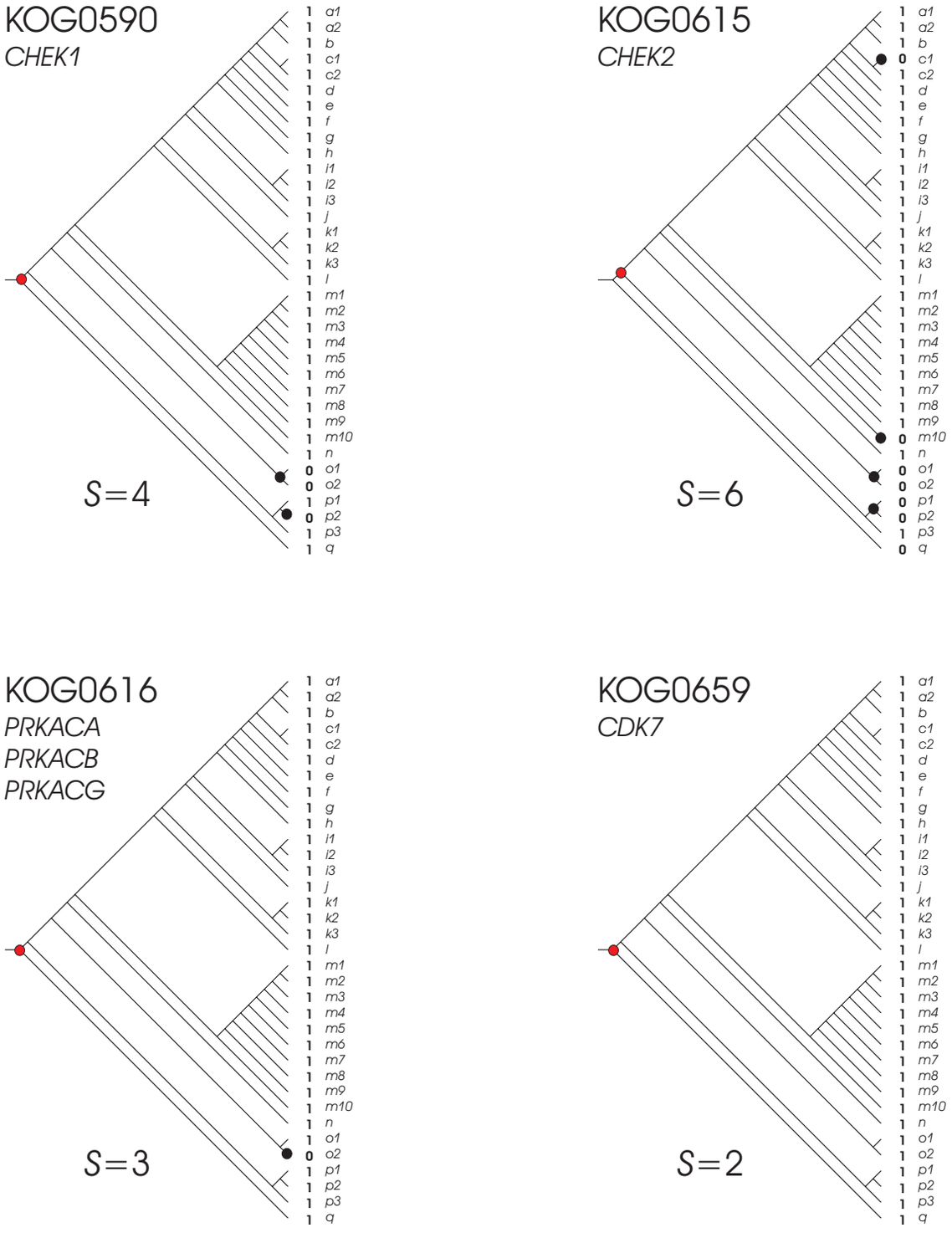
Supplementary Figure S16. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0218, KOG0219, KOG0220 and KOG0221, as in Supplementary Figure S14.



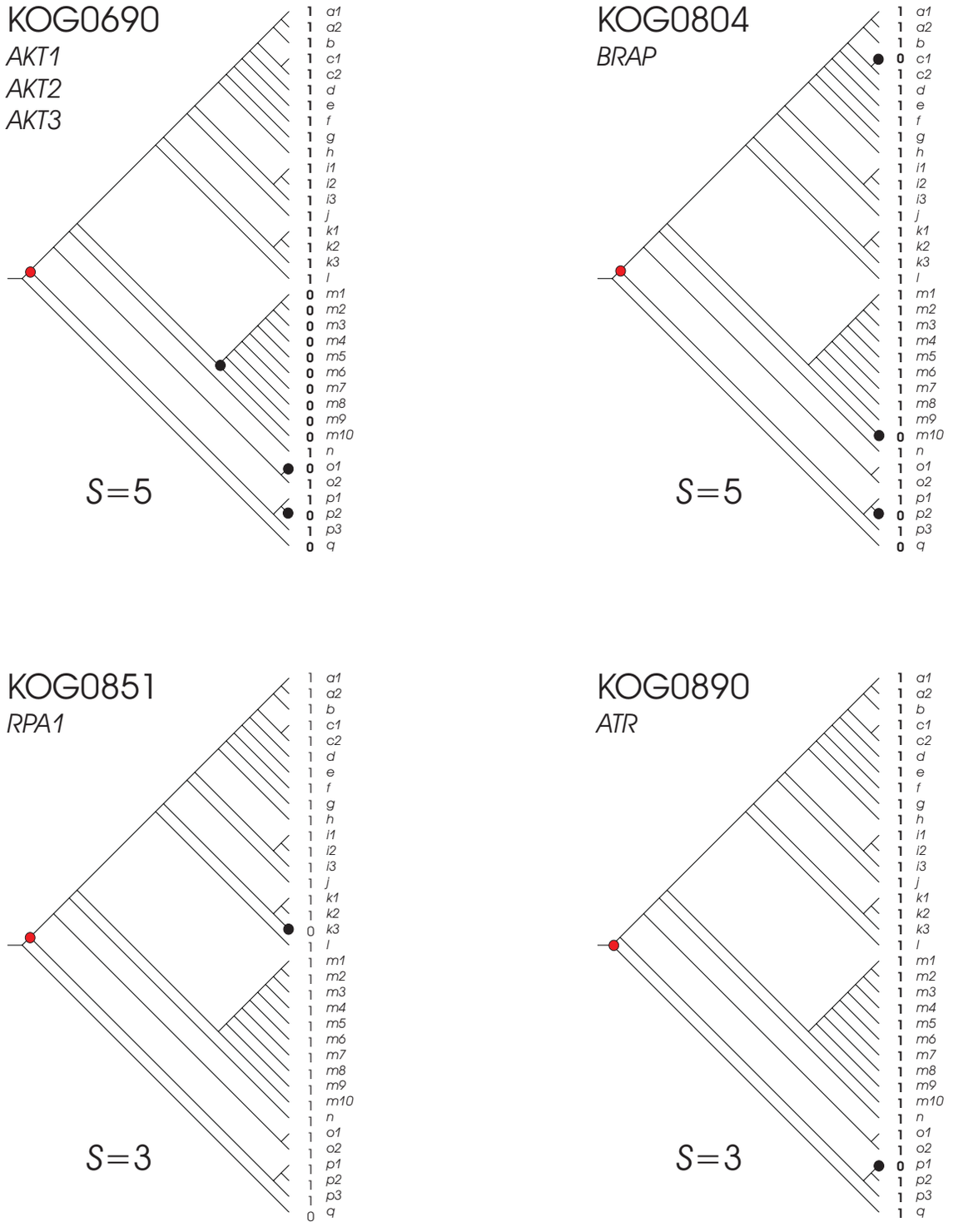
Supplementary Figure S17. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0297, KOG0375, KOG0381 and KOG0387, as in Supplementary Figure S14.



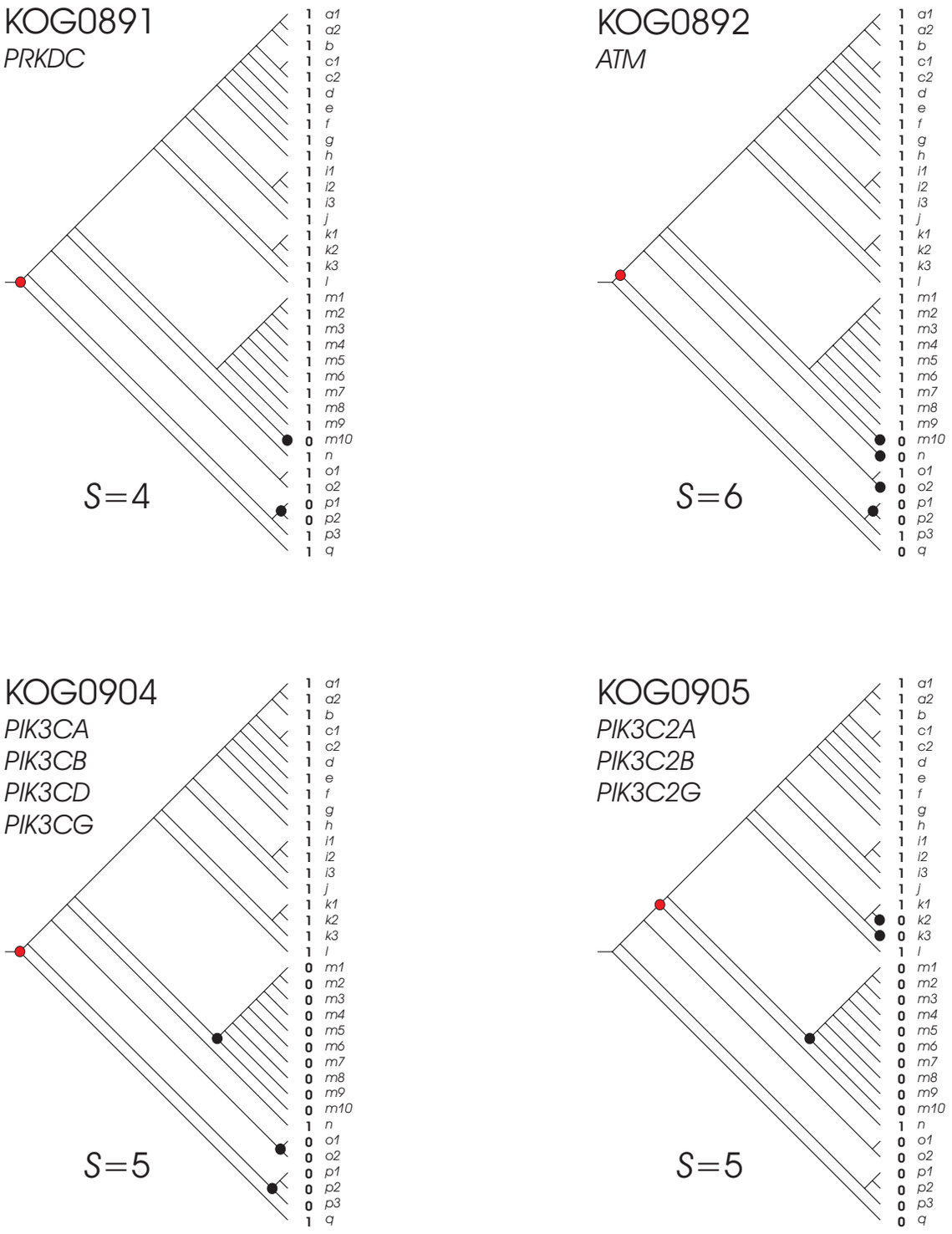
Supplementary Figure S18. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0390, KOG0442, KOG0504 and KOG0578, as in Supplementary Figure S14.



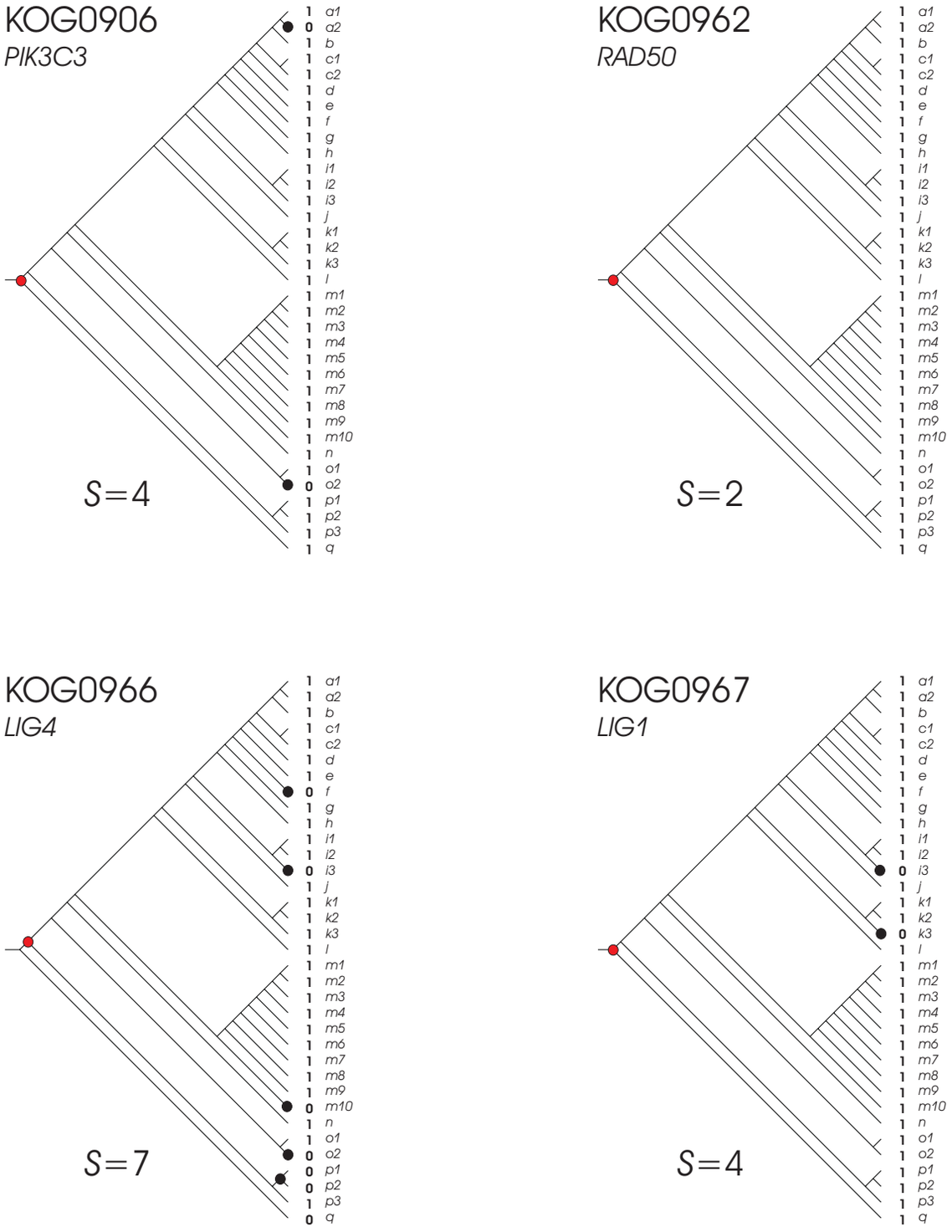
Supplementary Figure S19. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0590, KOG0615, KOG0616 and KOG0659, as in Supplementary Figure S14.



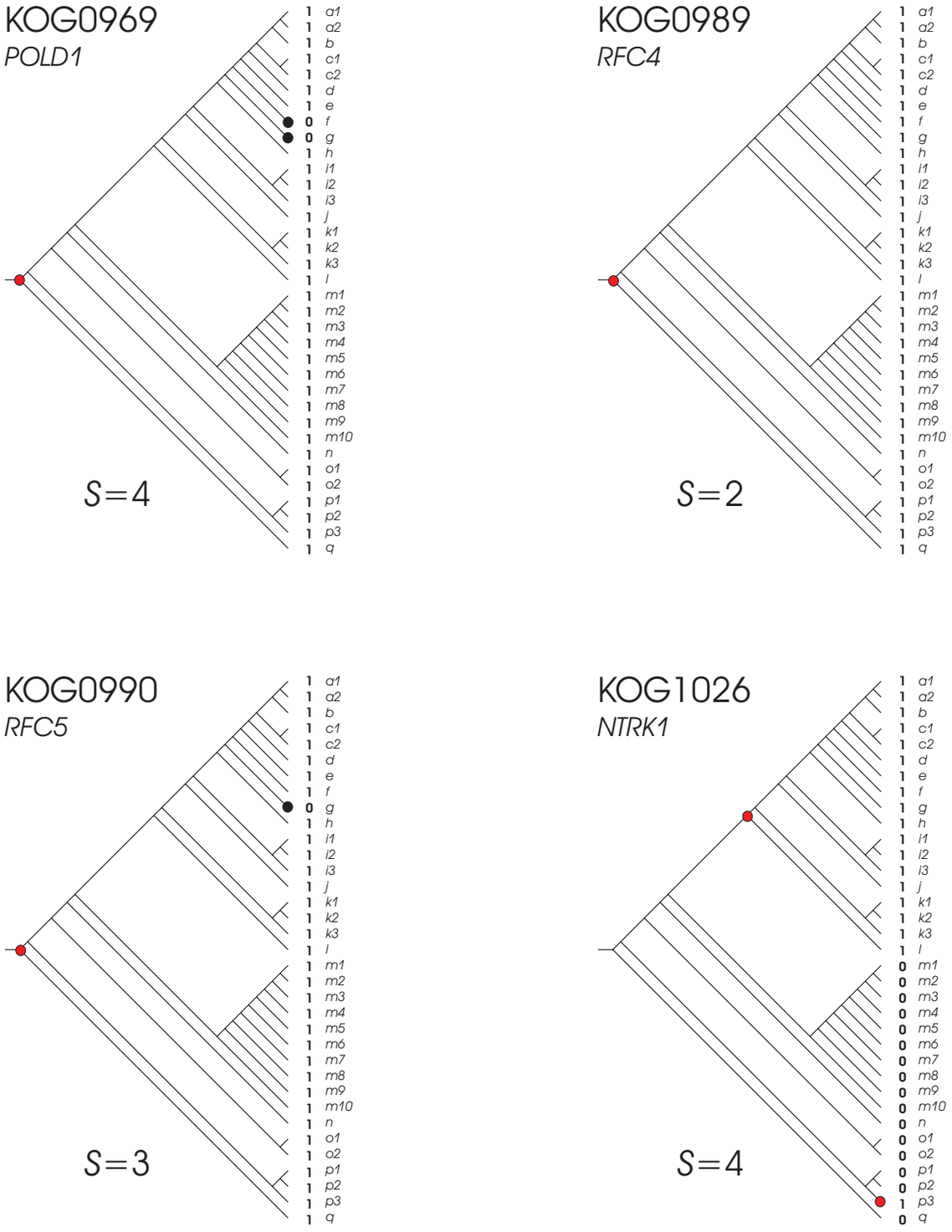
Supplementary Figure S20. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0690, KOG0804, KOG0851 and KOG0890, as in Supplementary Figure S14.



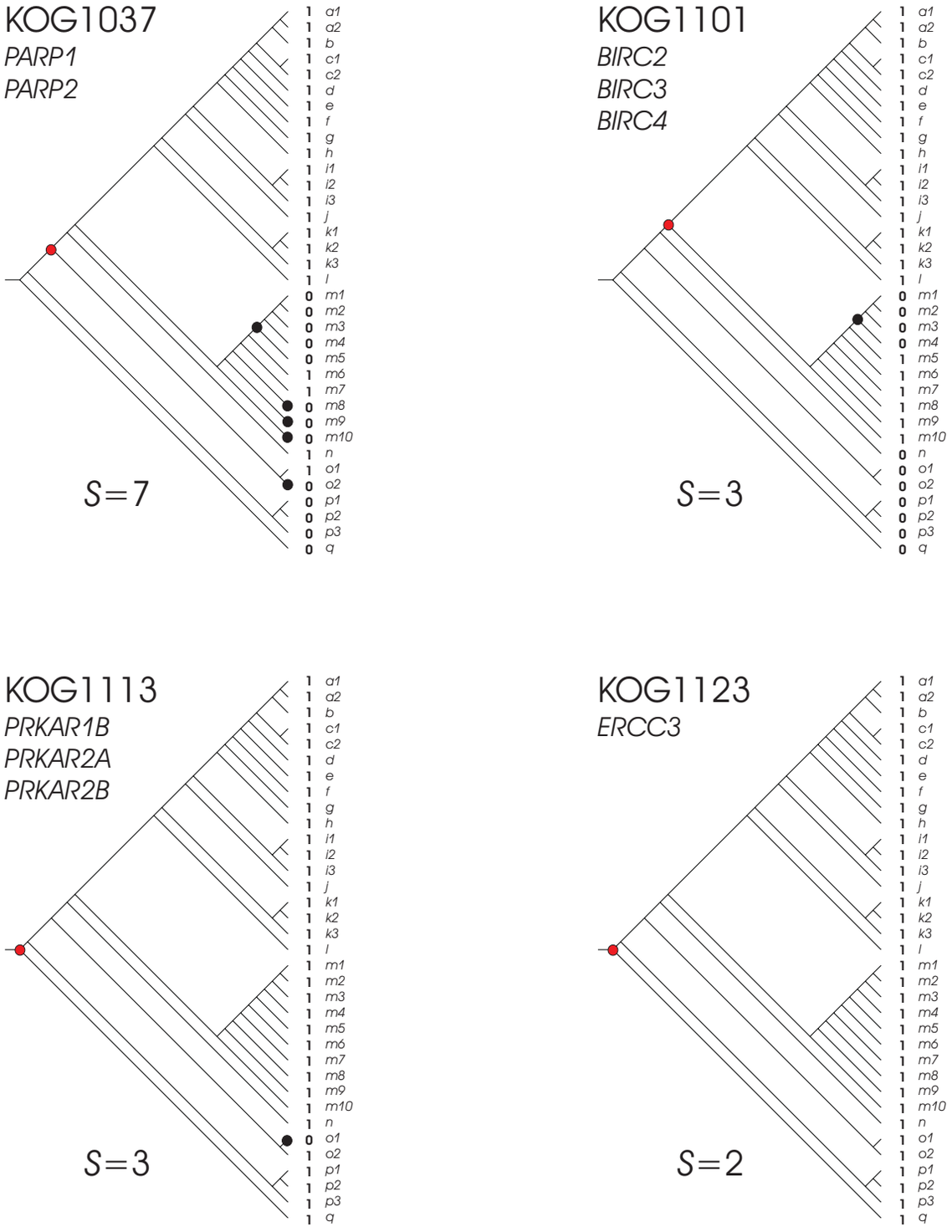
Supplementary Figure S21. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0891, KOG0892, KOG0904 and KOG0905, as in Supplementary Figure S14.



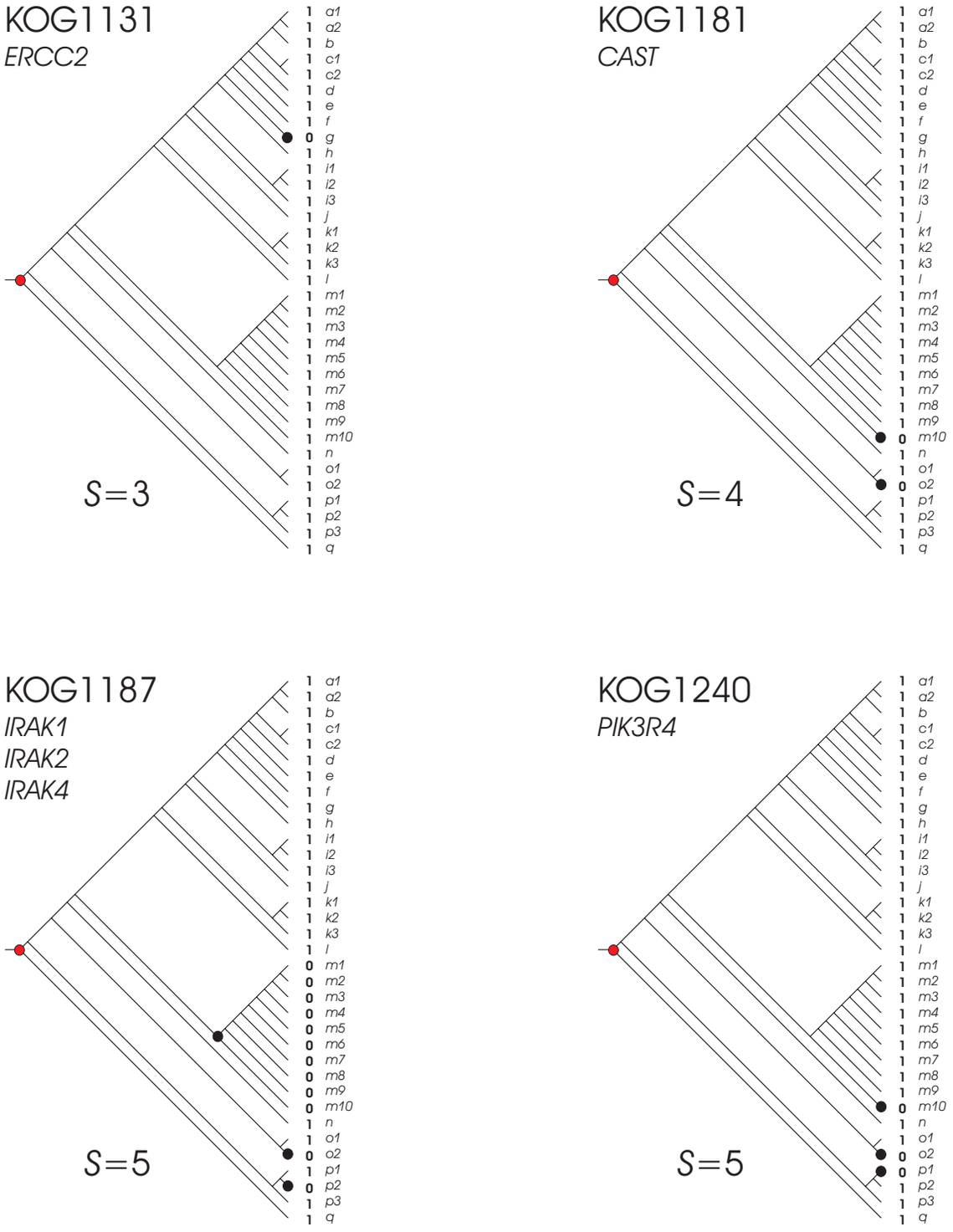
Supplementary Figure S22. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0906, KOG0962, KOG0966 and KOG0967, as in Supplementary Figure S14.



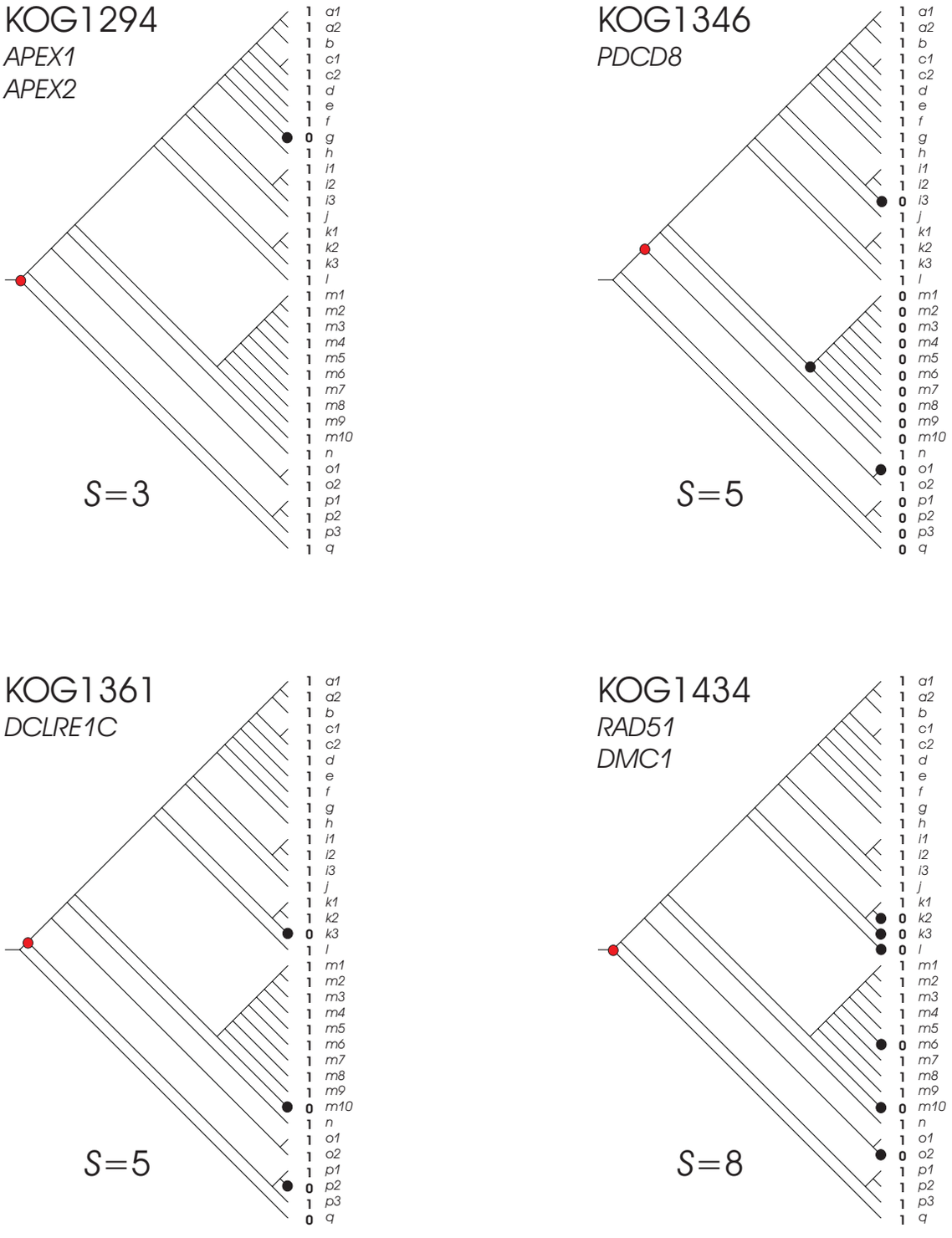
Supplementary Figure S23. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG0969, KOG0989, KOG0990 and KOG1026, as in Supplementary Figure S14.



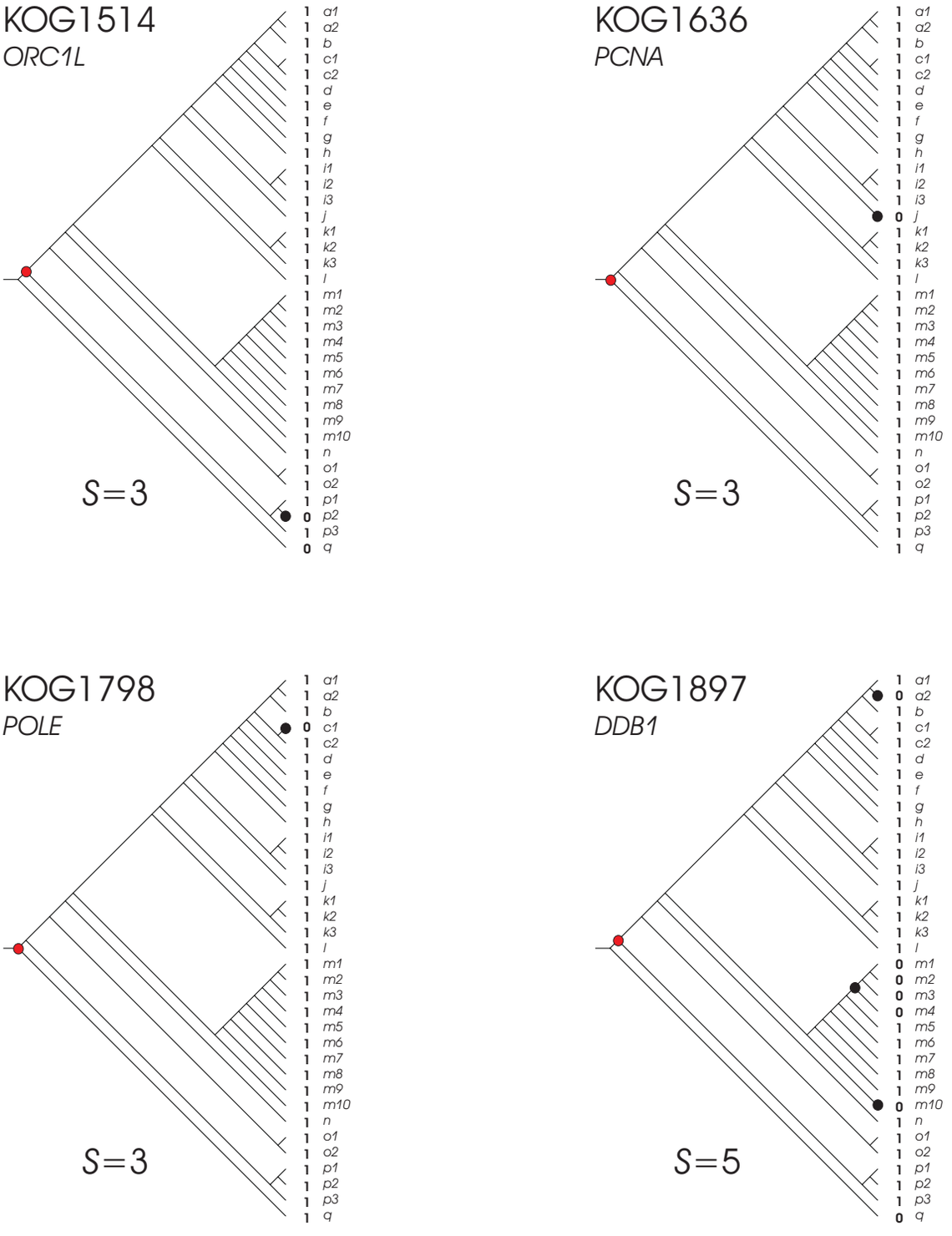
Supplementary Figure S24. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1037, KOG1101, KOG1113 and KOG1123, as in Supplementary Figure S14.



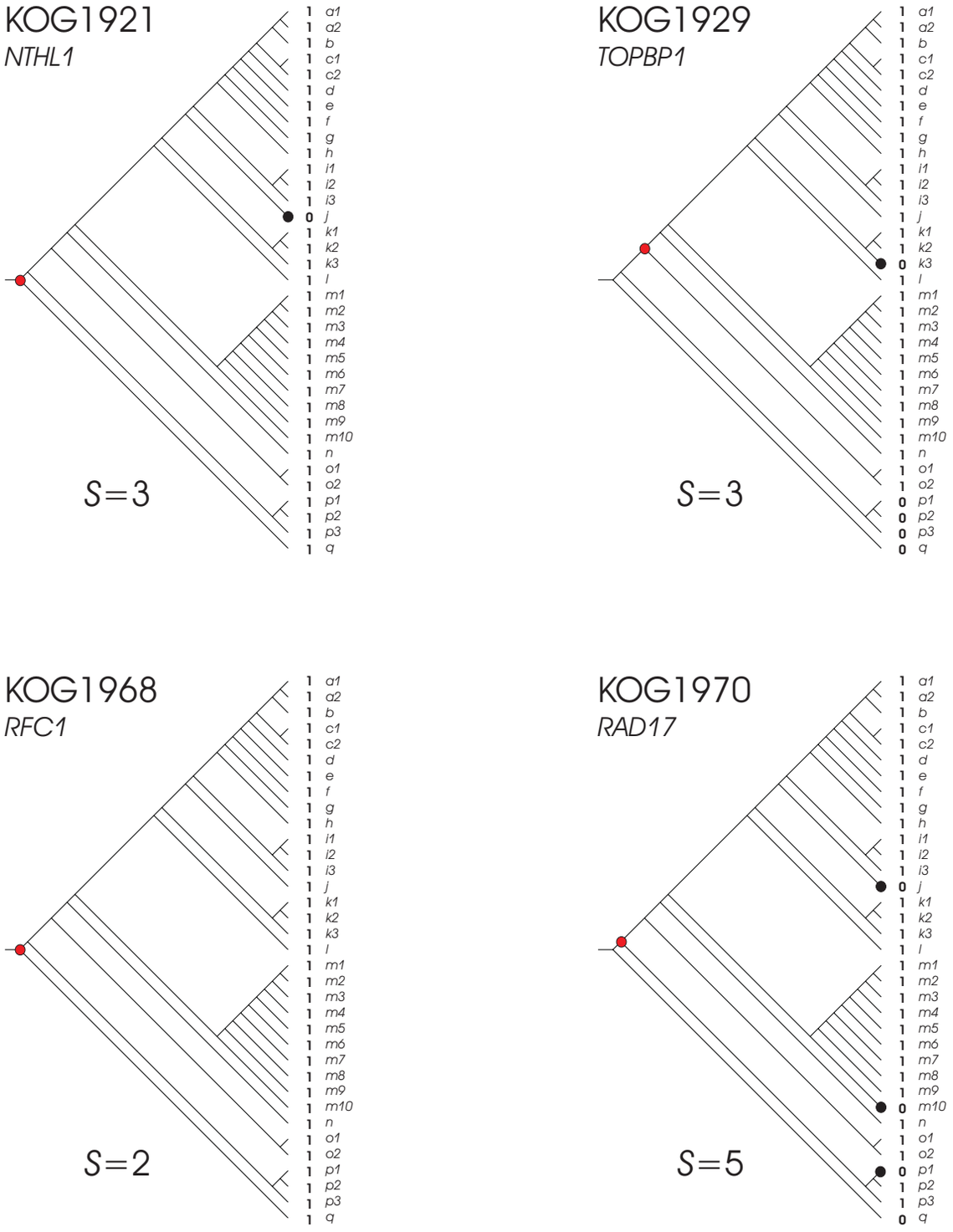
Supplementary Figure S25. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1131, KOG1181, KOG1187 and KOG1240, as in Supplementary Figure S14.



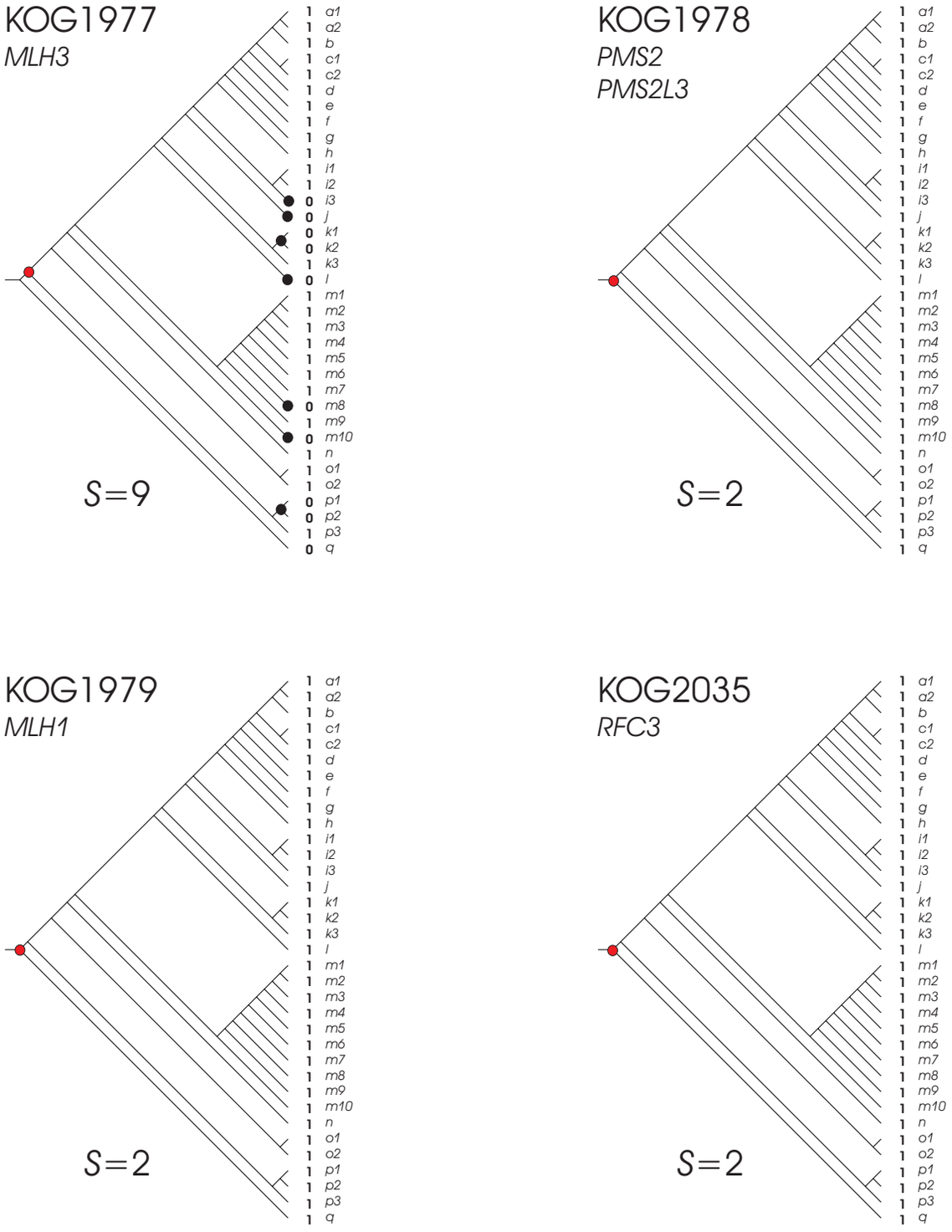
Supplementary Figure S26. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1294, KOG1346, KOG1361 and KOG1434, as in Supplementary Figure S14.



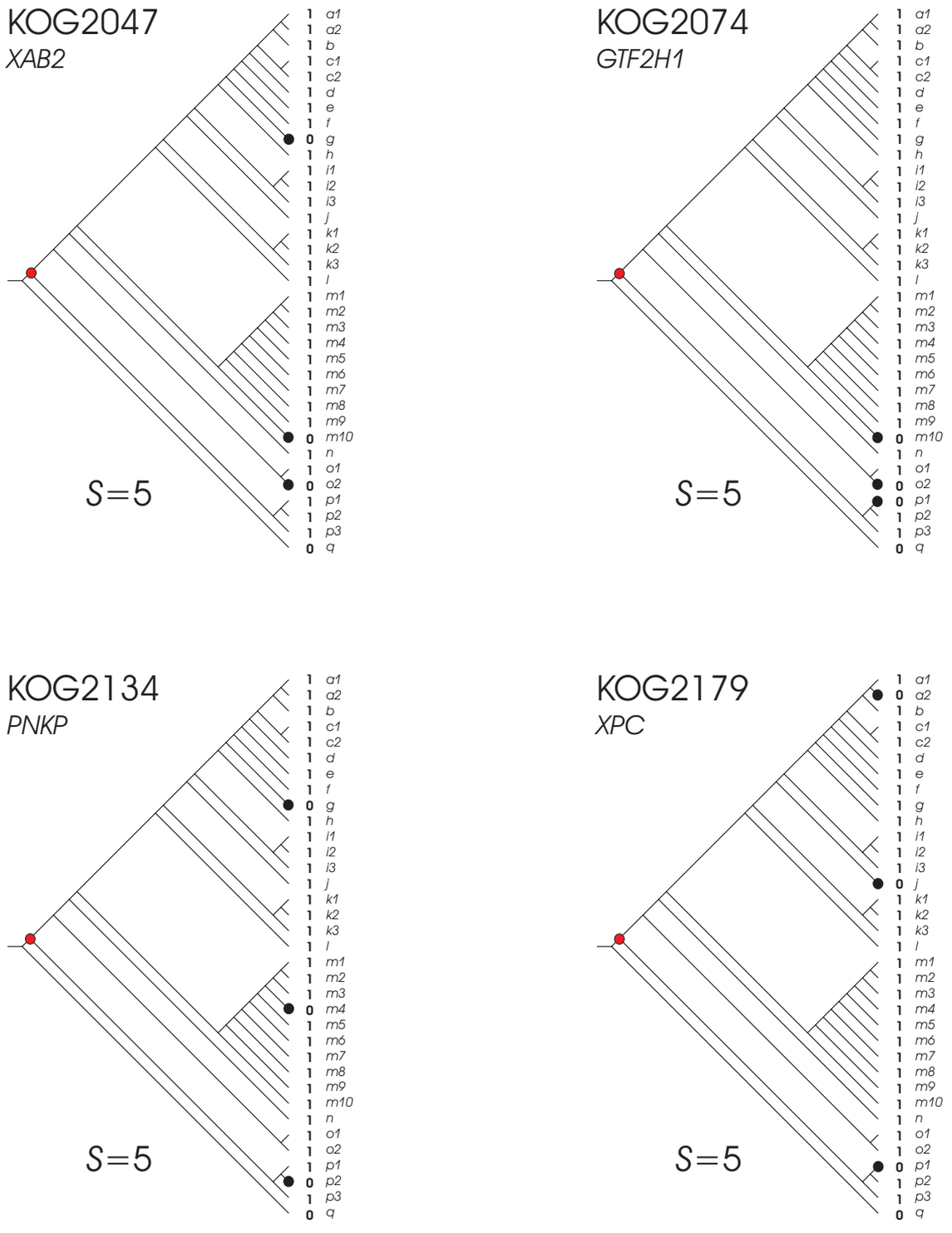
Supplementary Figure S27. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1514, KOG1636, KOG1798 and KOG1897, as in Supplementary Figure S14.



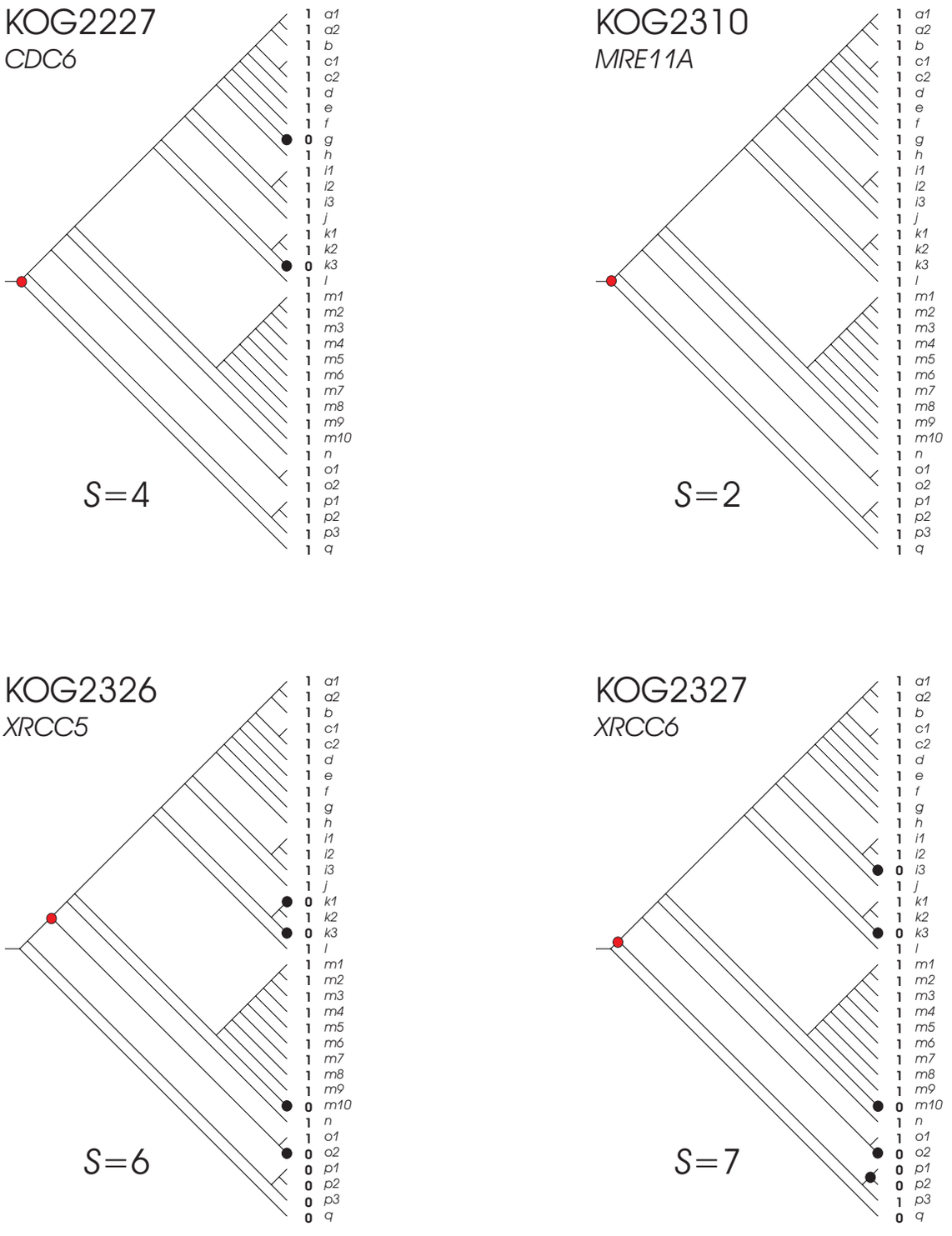
Supplementary Figure S28. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1921, KOG1929, KOG1968 and KOG1970, as in Supplementary Figure S14.



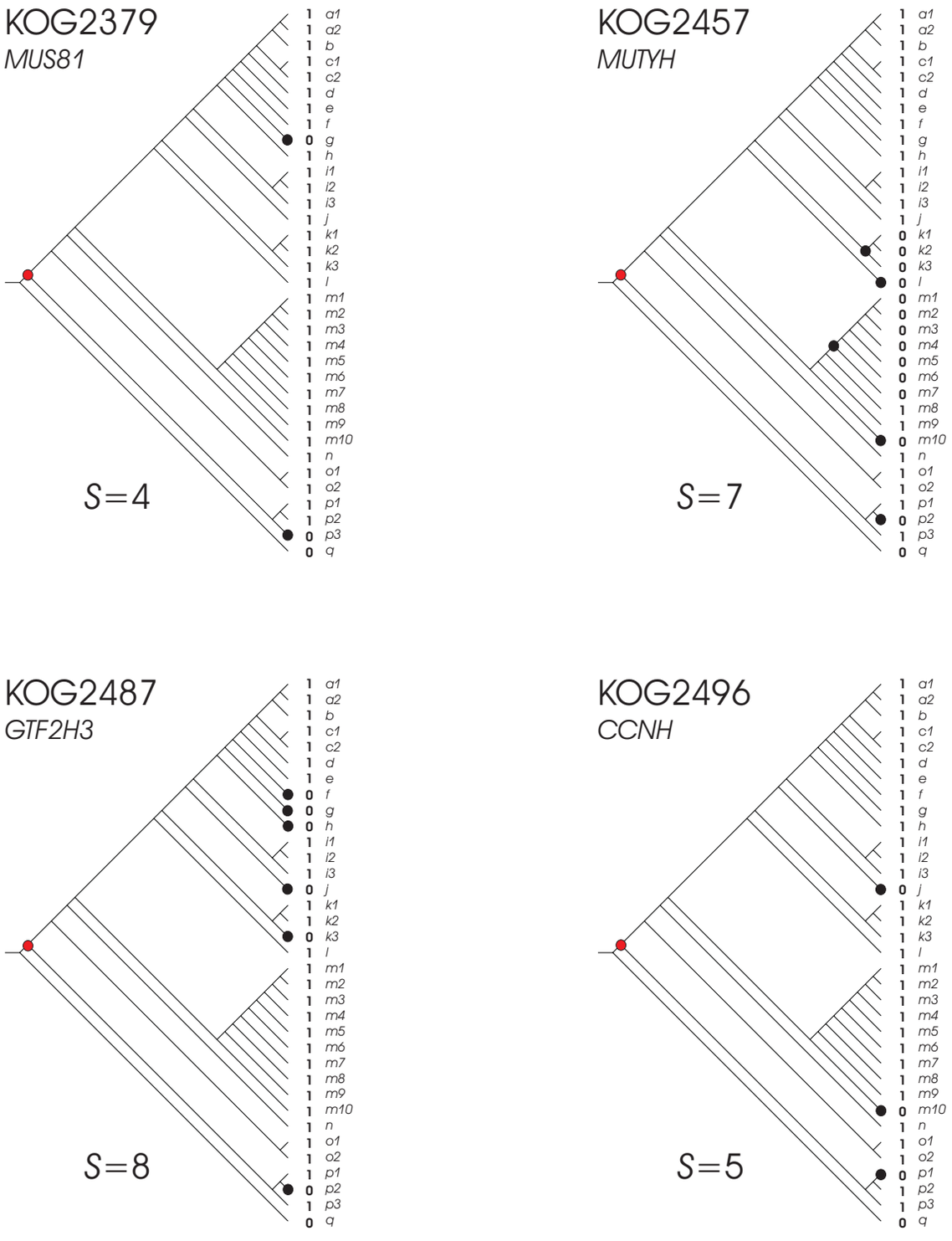
Supplementary Figure S29. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG1977, KOG1978, KOG1979 and KOG2035, as in Supplementary Figure S14.



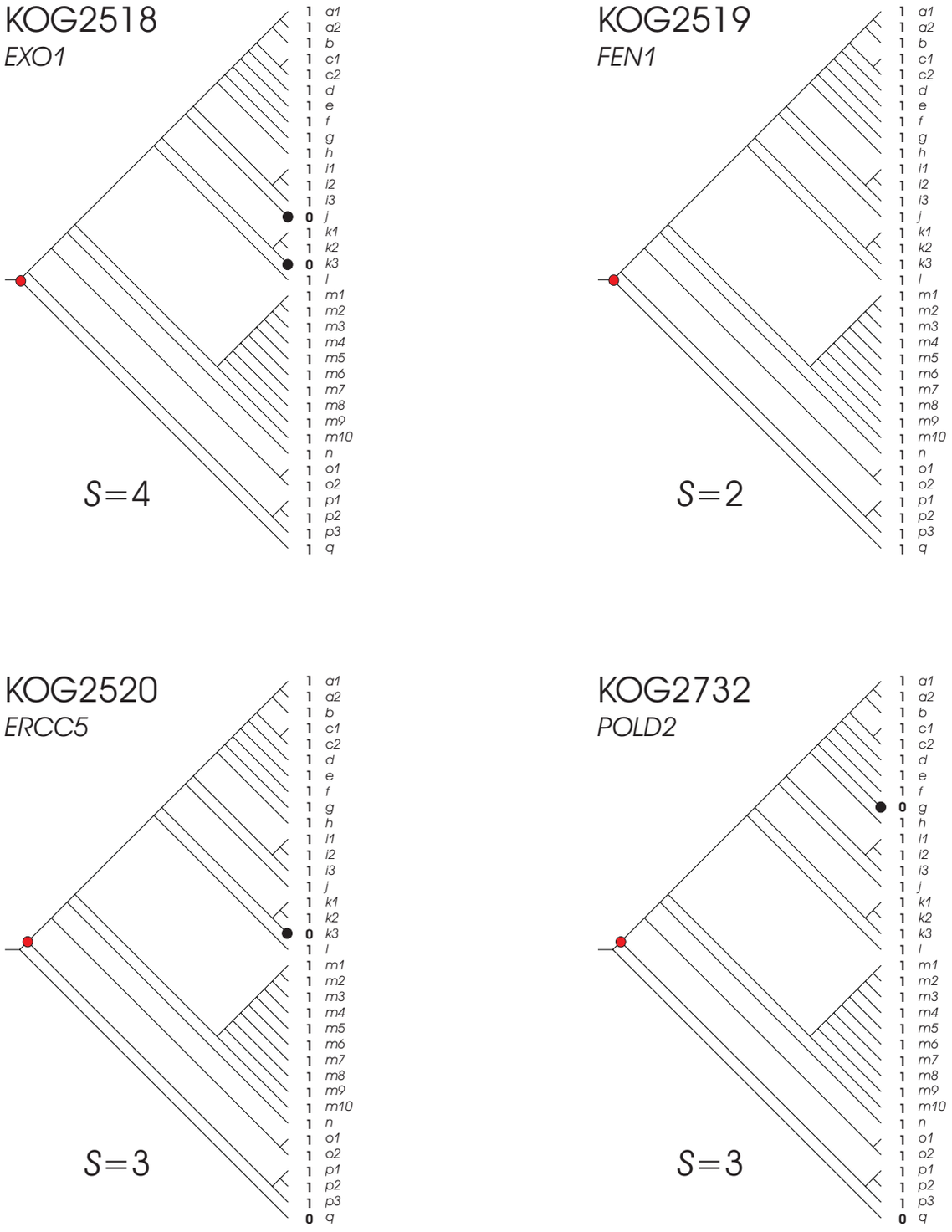
Supplementary Figure S30. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2047, KOG2074, KOG2134 and KOG2179, as in Supplementary Figure S14.



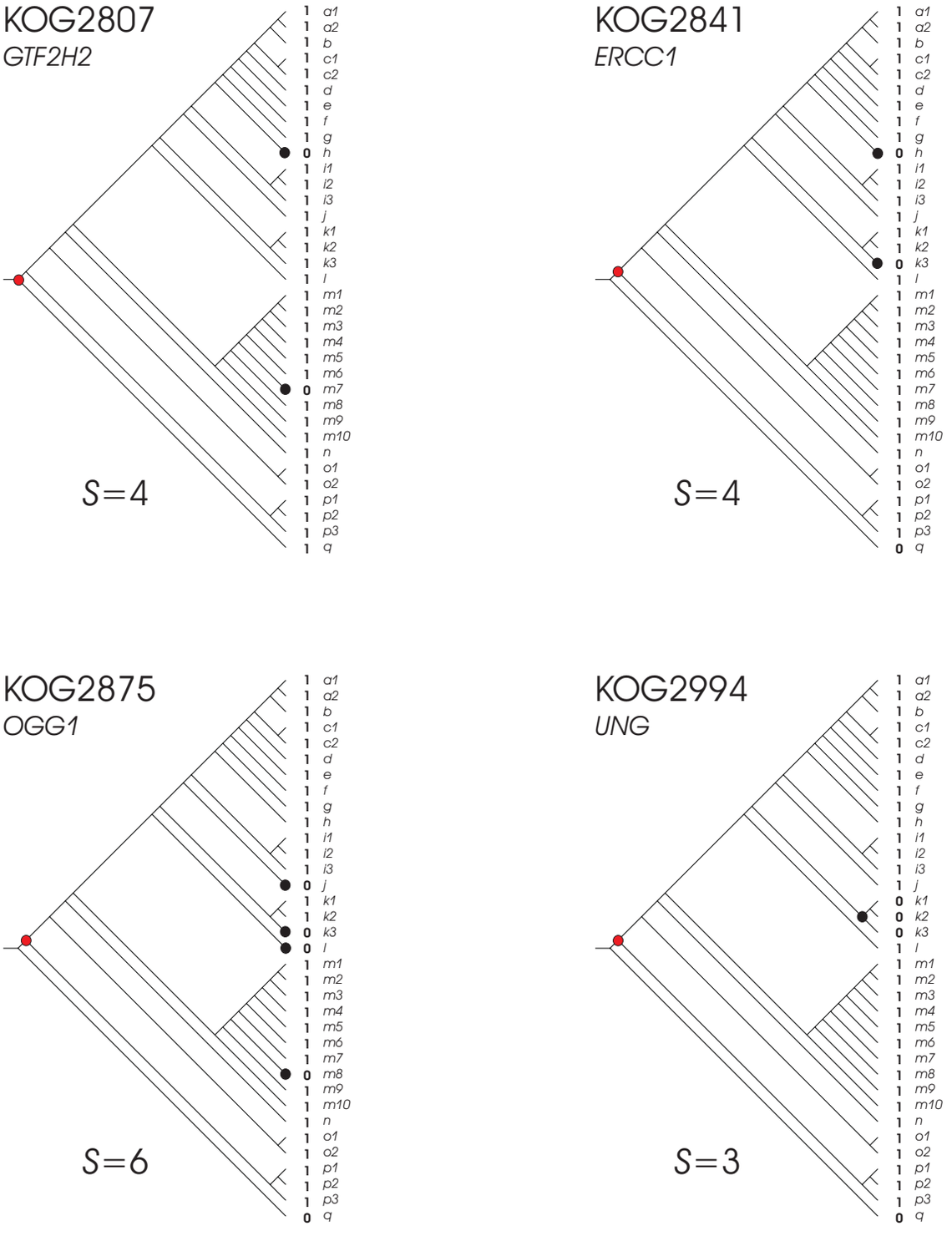
Supplementary Figure S31. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2227, KOG2310, KOG2326 and KOG2327, as in Supplementary Figure S14.



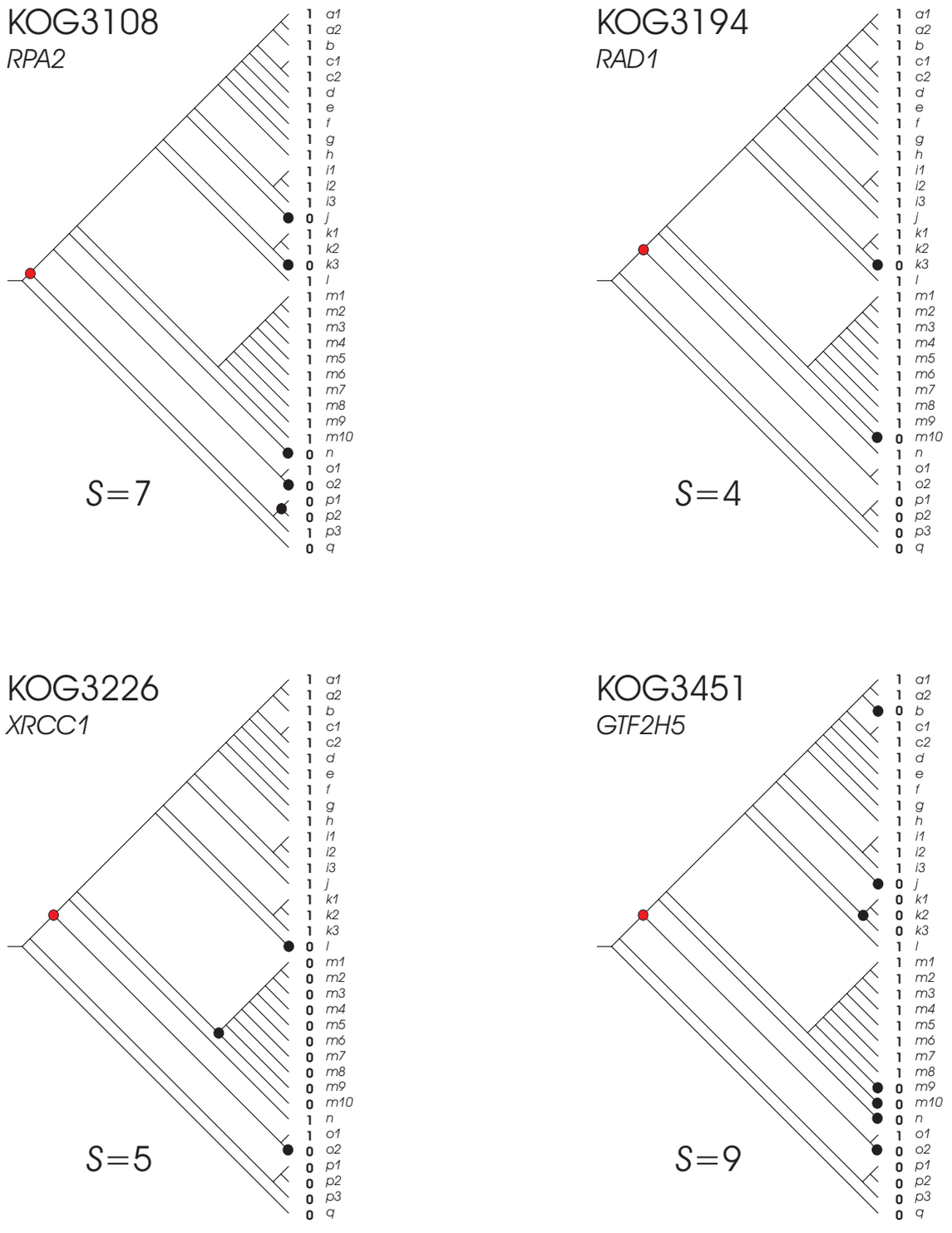
Supplementary Figure S32. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2379, KOG2457, KOG2487 and KOG2496, as in Supplementary Figure S14.



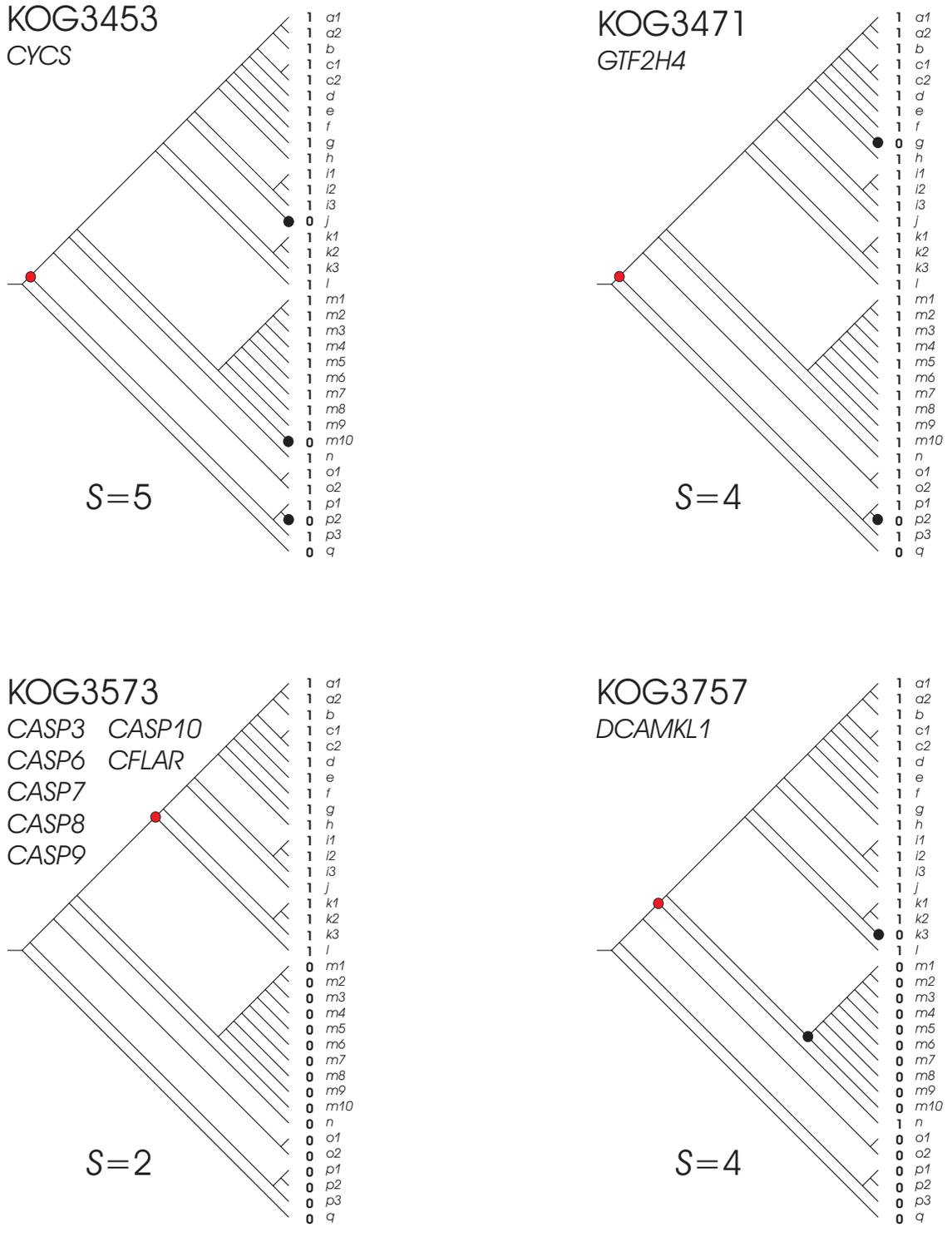
Supplementary Figure S33. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2518, KOG2519, KOG2520 and KOG2732, as in Supplementary Figure S14.



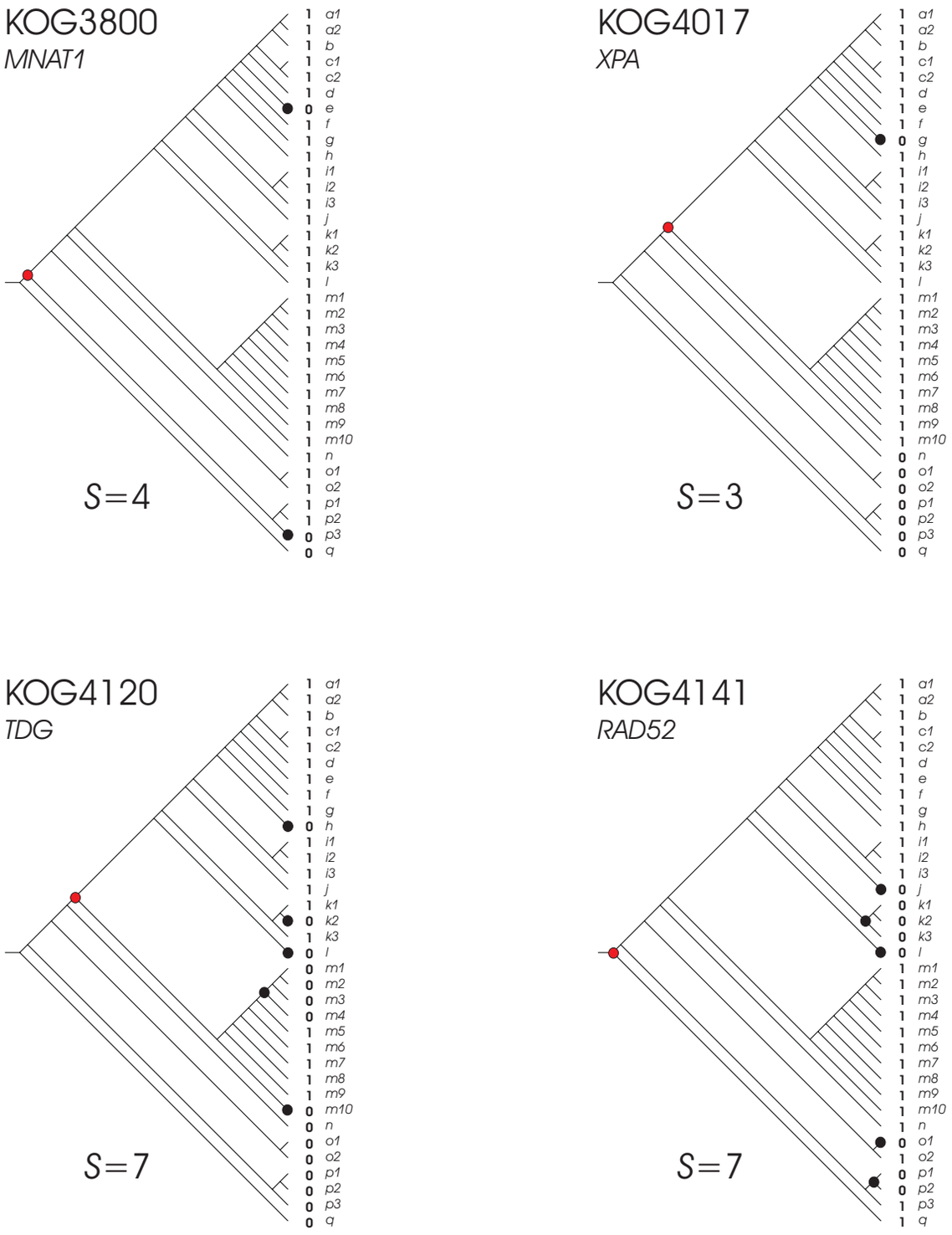
Supplementary Figure S34. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG2807, KOG2841, KOG2875 and KOG2994, as in Supplementary Figure S14.



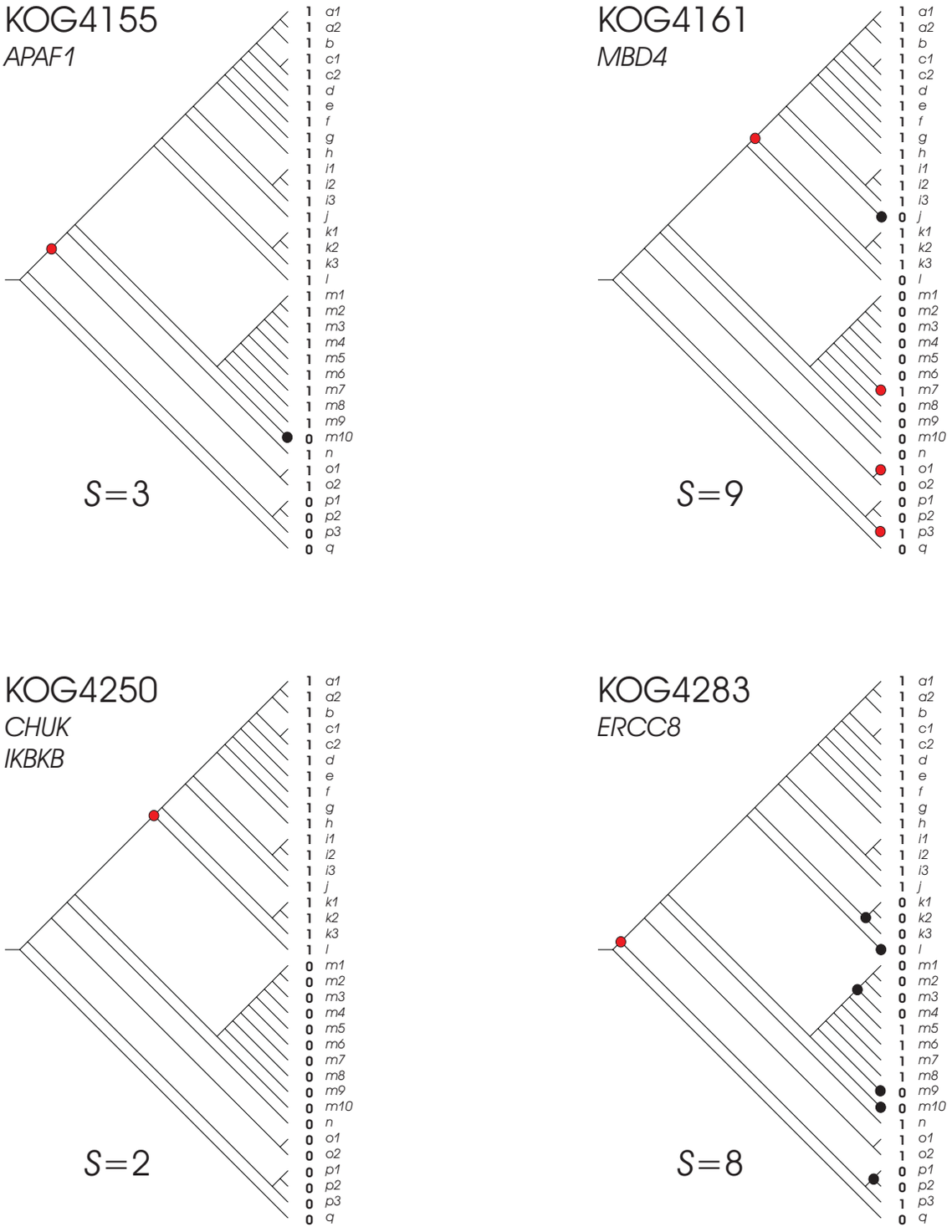
Supplementary Figure S35. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3108, KOG3194, KOG3226 and KOG3451, as in Supplementary Figure S14.



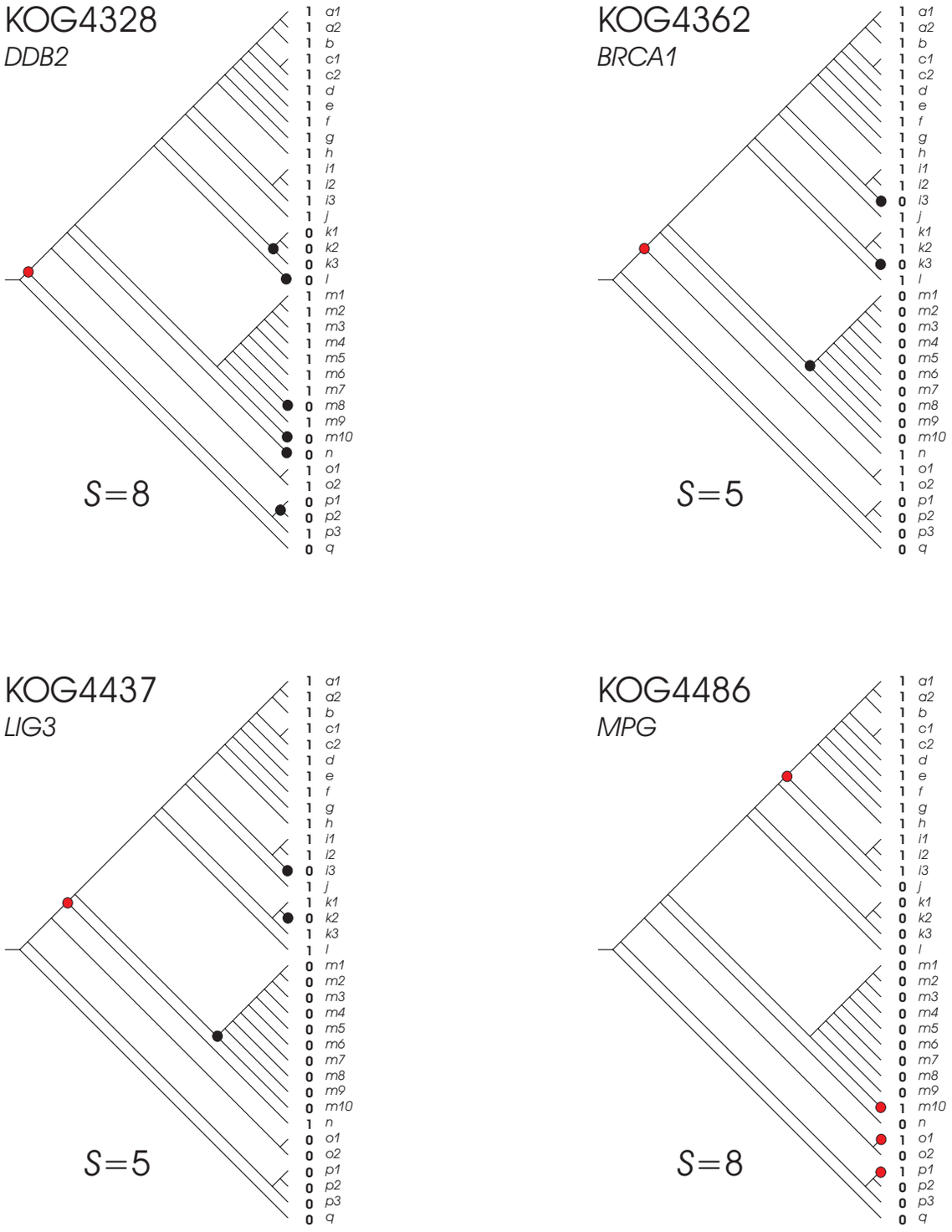
Supplementary Figure S36. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3453, KOG3471, KOG3573 and KOG3757, as in Supplementary Figure S14.



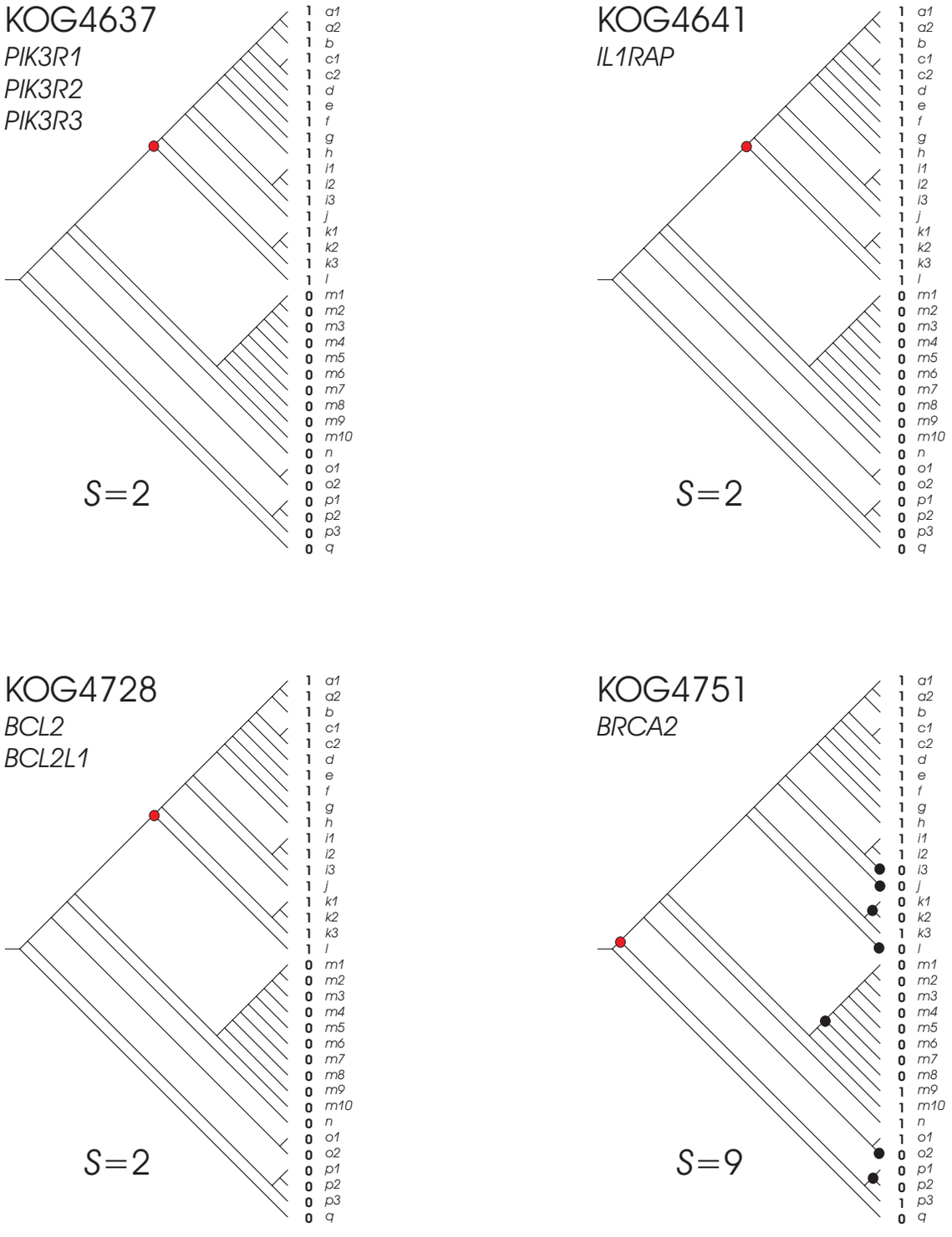
Supplementary Figure S37. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG3800, KOG4017, KOG4120 and KOG4141, as in Supplementary Figure S14.



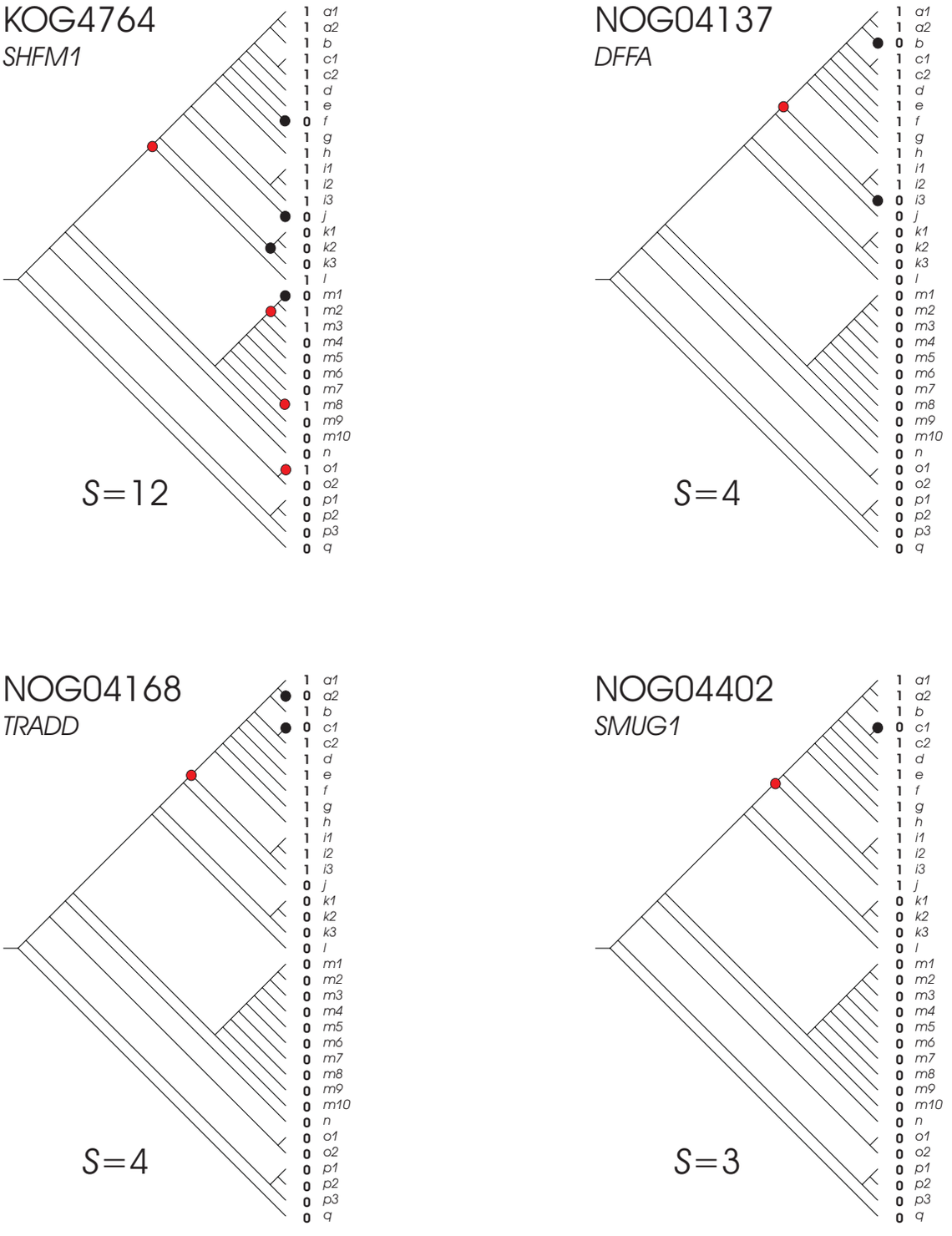
Supplementary Figure S38. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4155, KOG4161, KOG4250 and KOG4283, as in Supplementary Figure S14.



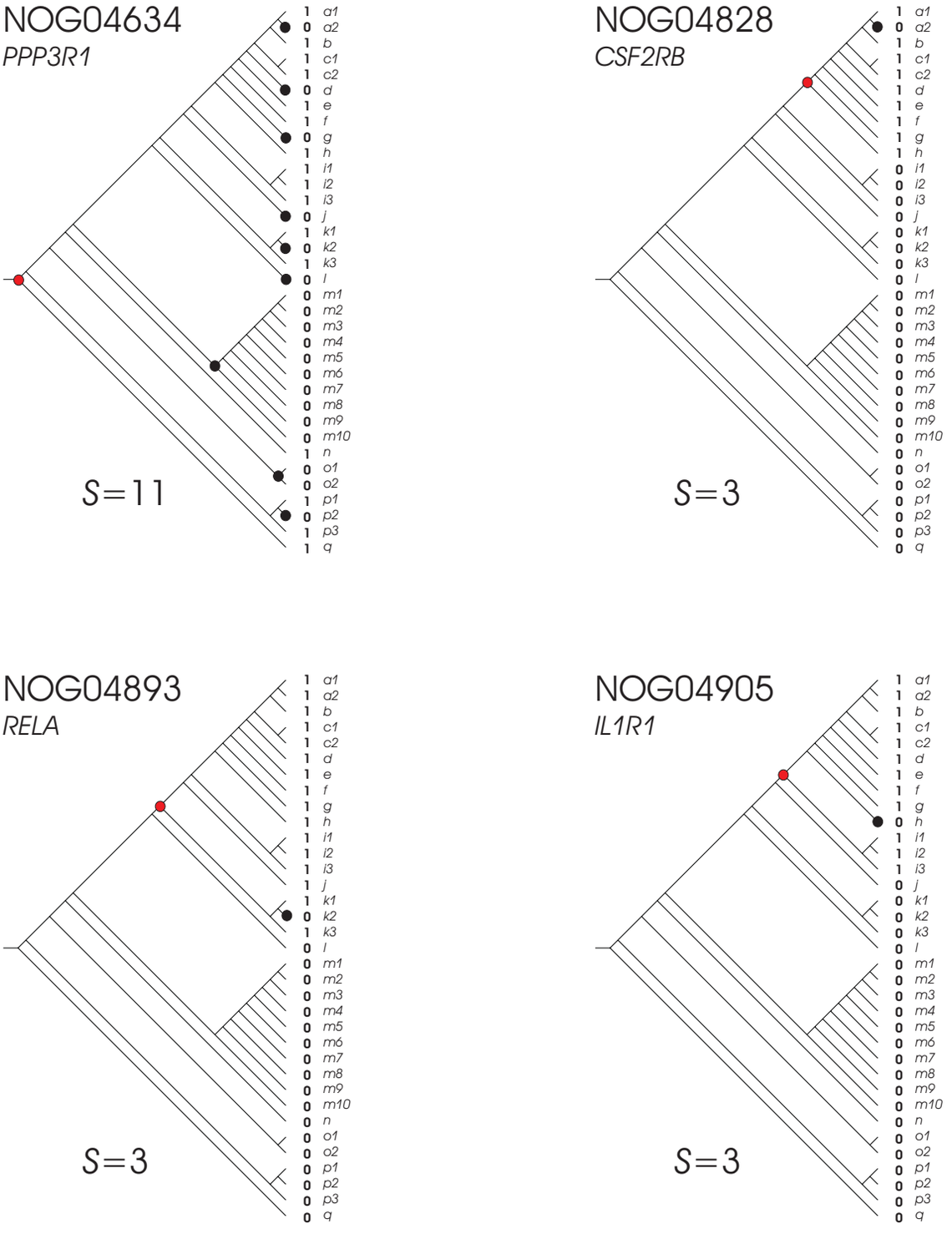
Supplementary Figure S39. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4328, KOG4362, KOG4437 and KOG4486, as in Supplementary Figure S14.



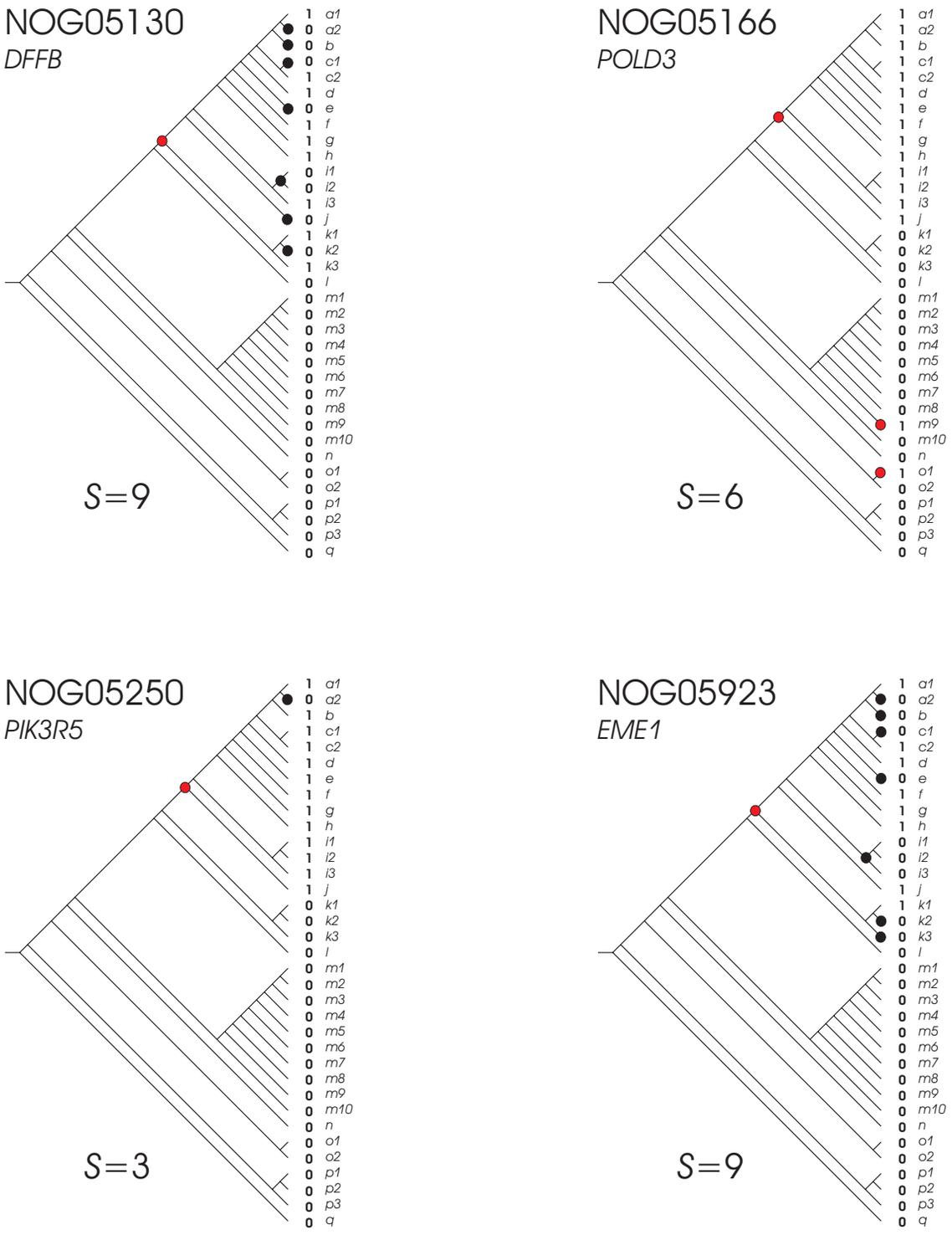
Supplementary Figure S40. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4637, KOG4641, KOG4728 and KOG4751, as in Supplementary Figure S14.



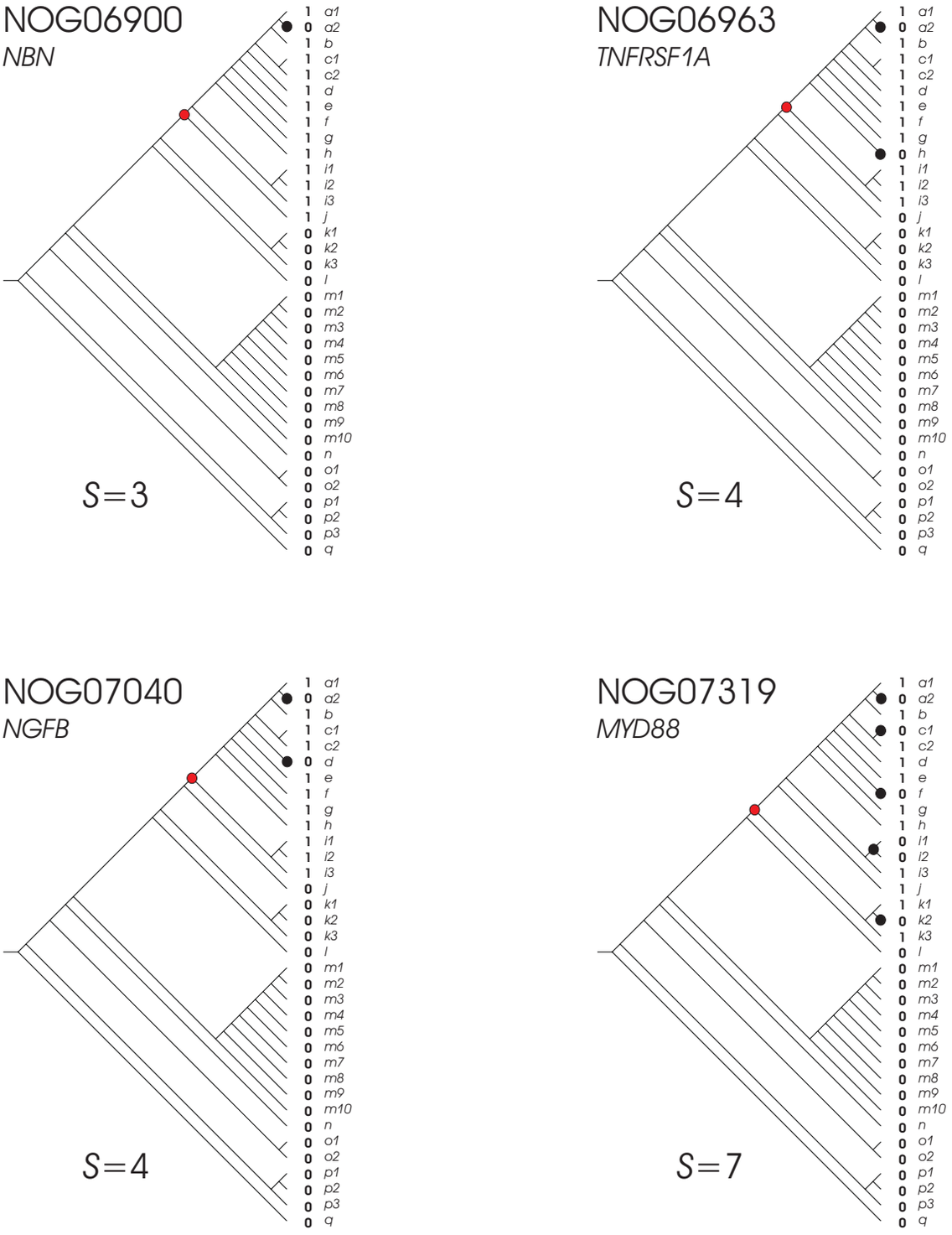
Supplementary Figure S41. Parsimony analysis of eukaryotic clusters of orthologous groups: KOG4764, NOG04137, NOG04168 and NOG04402, as in Supplementary Figure S14.



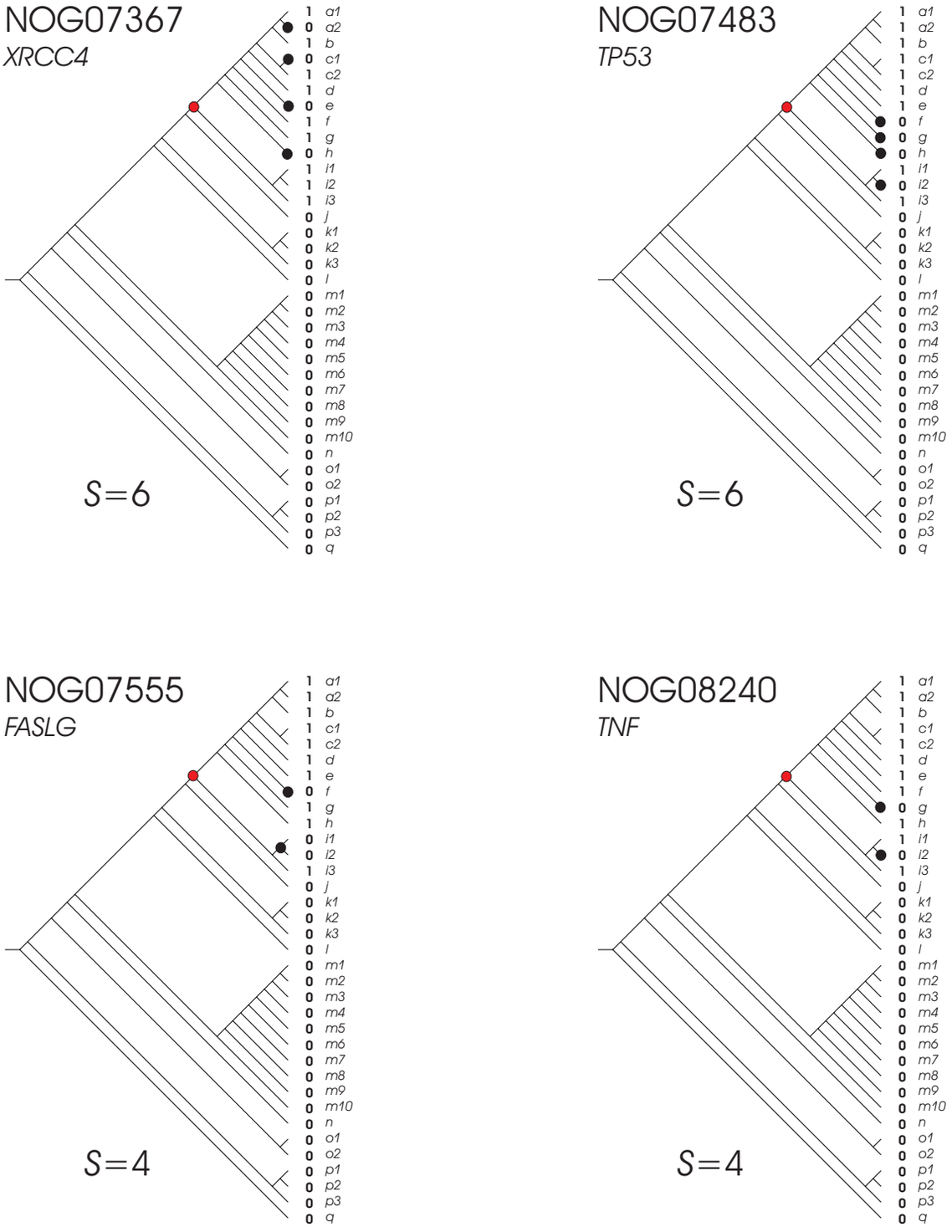
Supplementary Figure S42. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG04634, NOG04828, NOG04893 and NOG04905, as in Supplementary Figure S14.



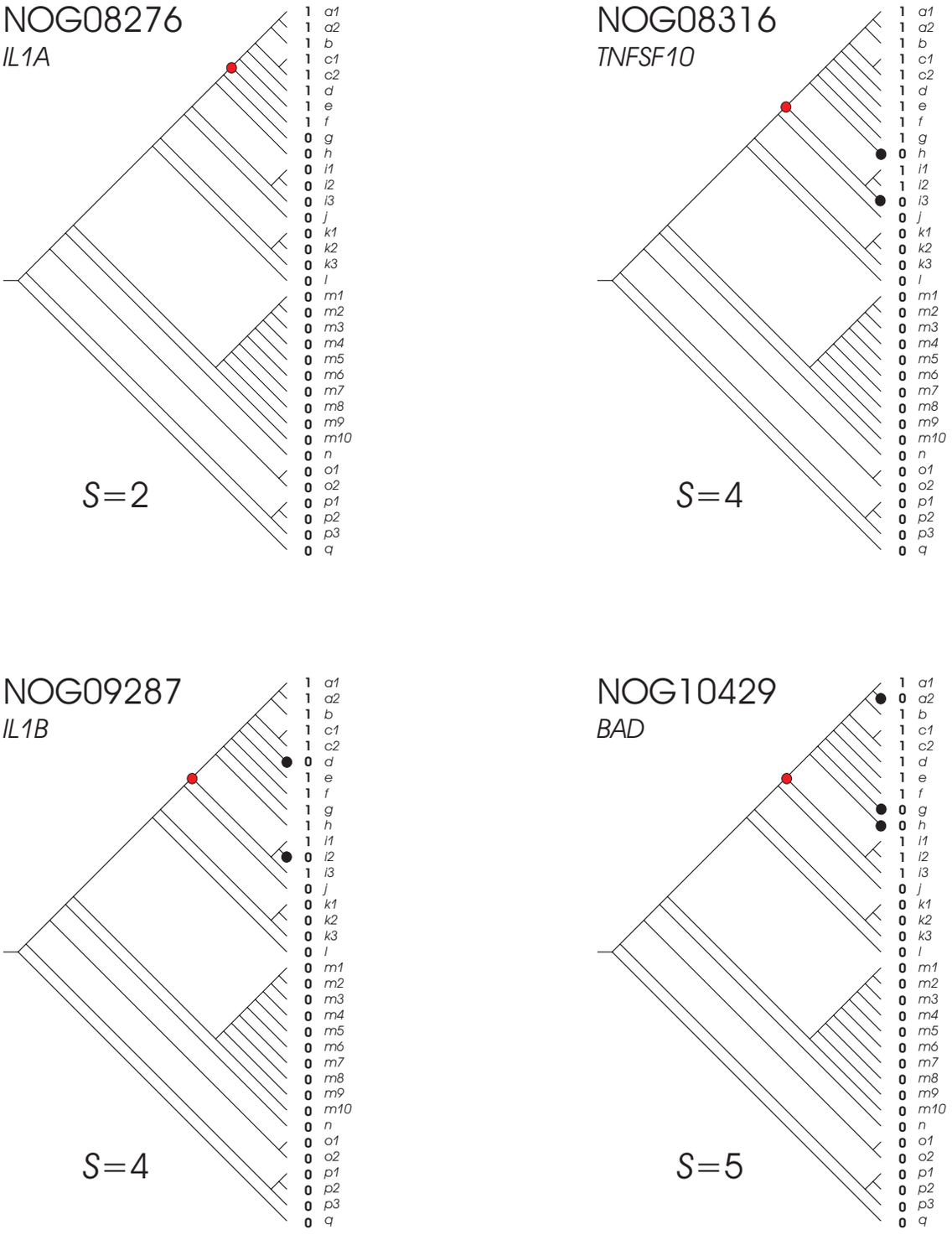
Supplementary Figure S43. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG05130, NOG05166, NOG05250 and NOG05923, as in Supplementary Figure S14.



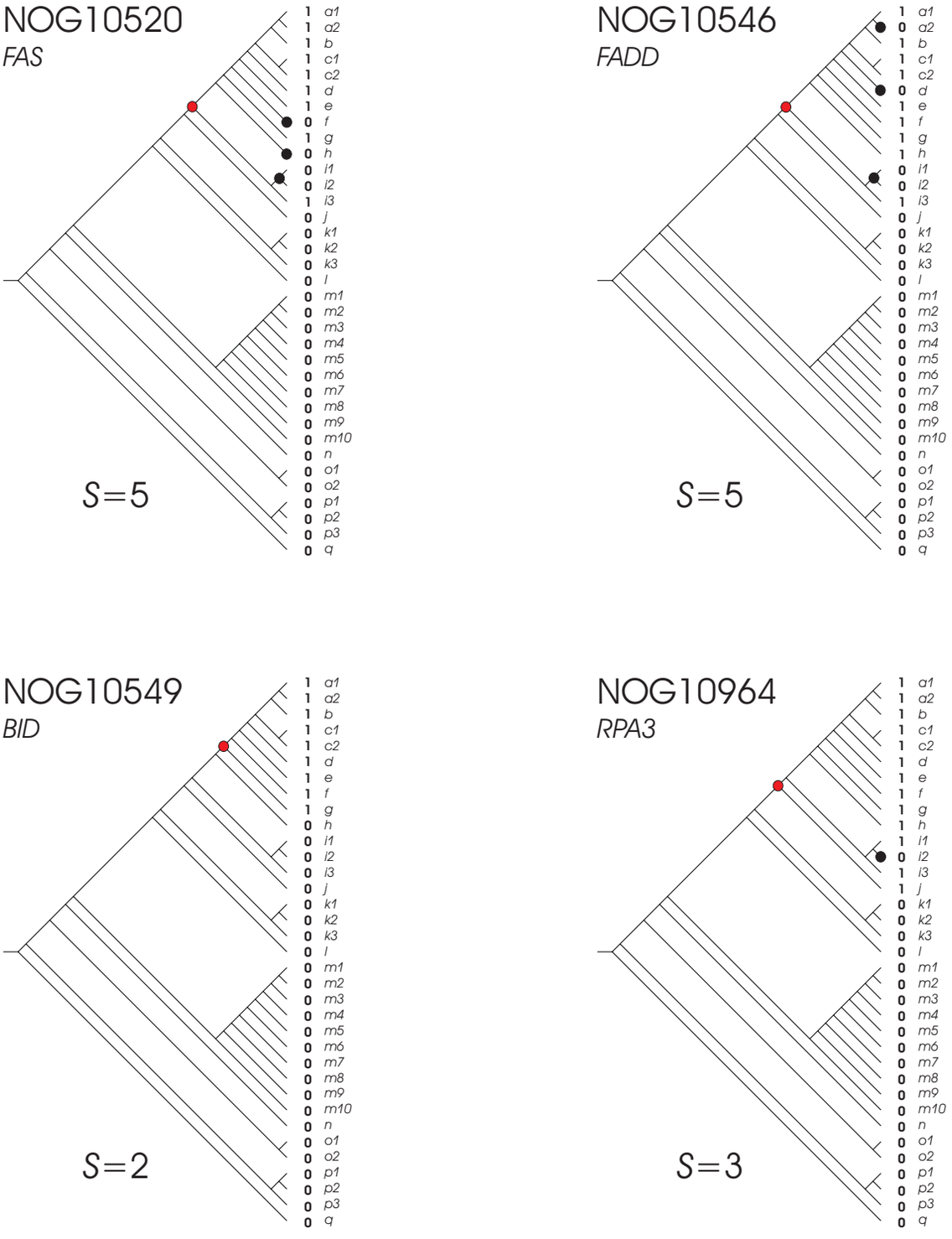
Supplementary Figure S44. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG06900, NOG06963, NOG07040 and NOG07319, as in Supplementary Figure S14.



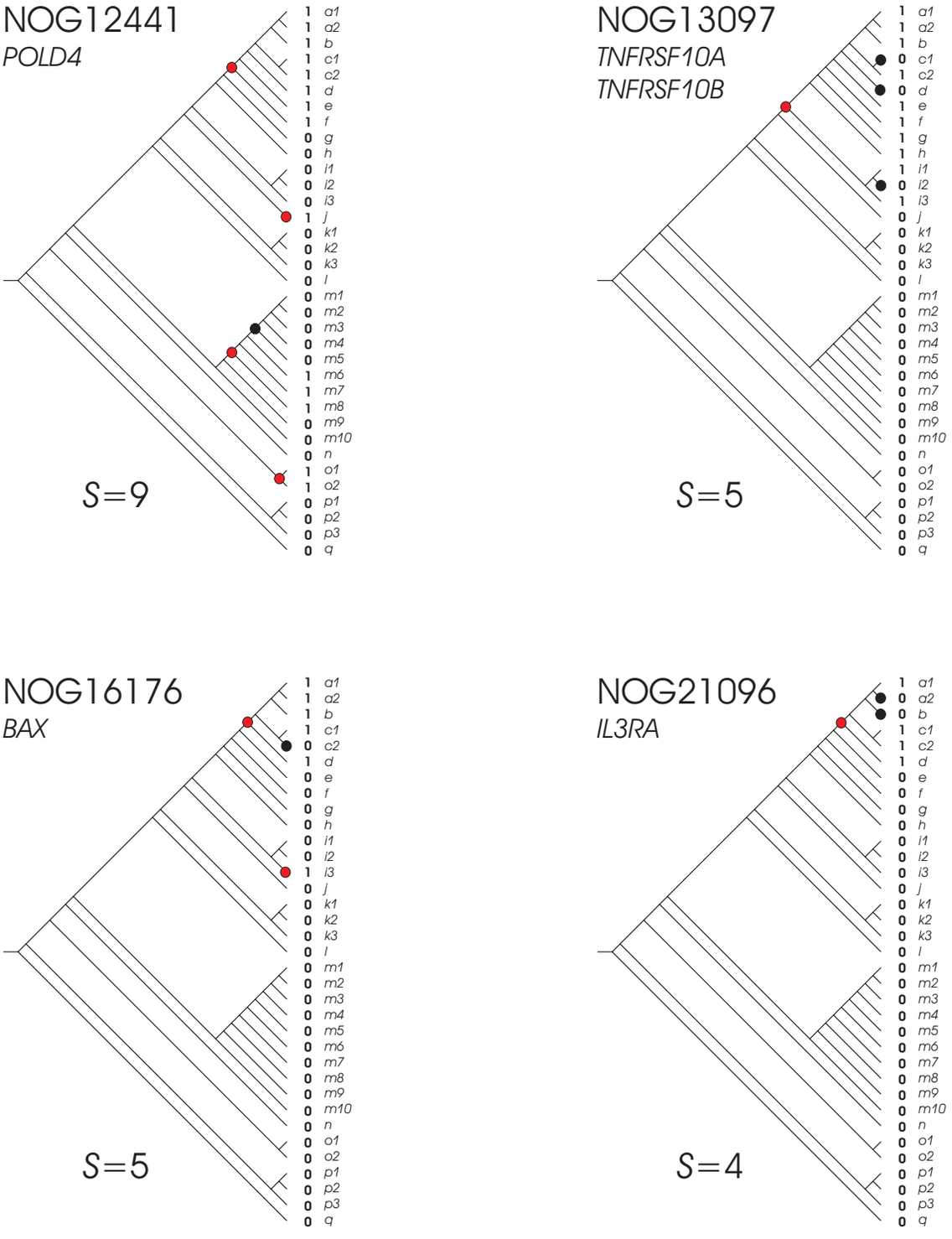
Supplementary Figure S45. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG07367, NOG07483, NOG07555 and NOG08240, as in Supplementary Figure S14.



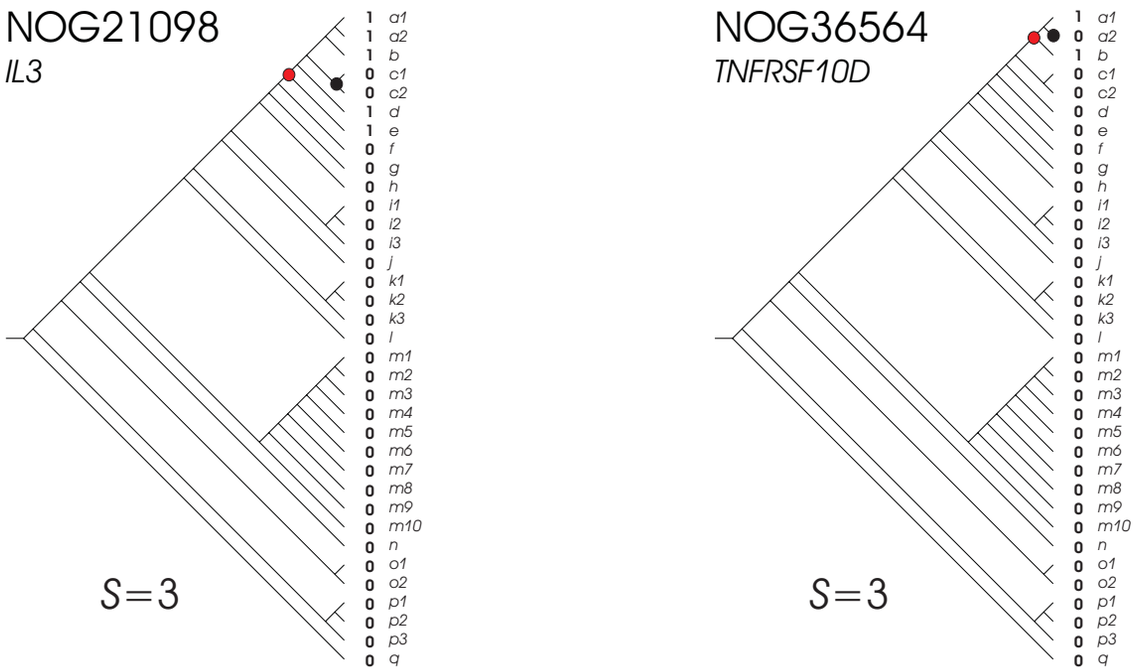
Supplementary Figure S46. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG08276, NOG08316, NOG09287 and NOG10429, as in Supplementary Figure S14.



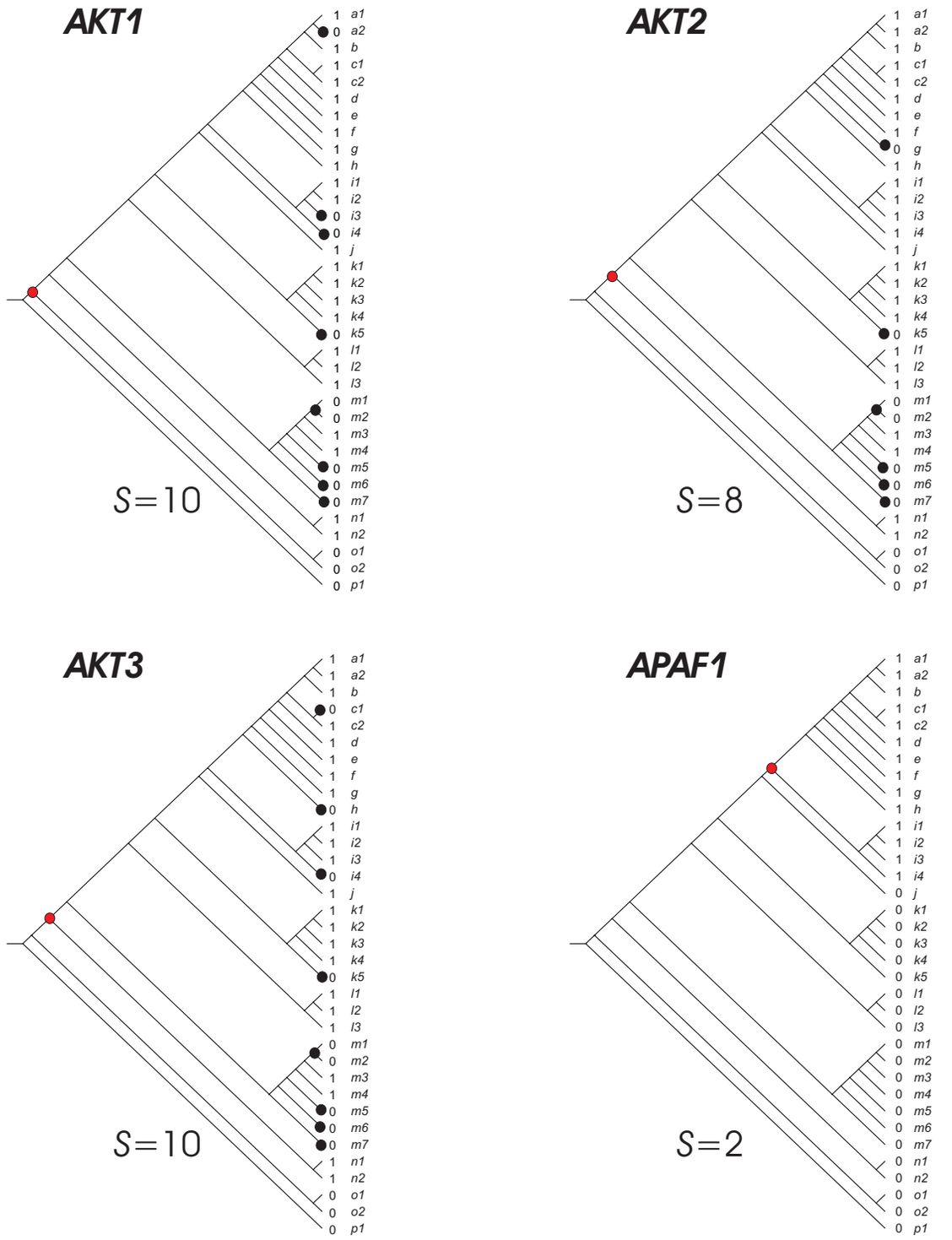
Supplementary Figure S47. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG10520, NOG10546, NOG10549 and NOG10964, as in Supplementary Figure S14.



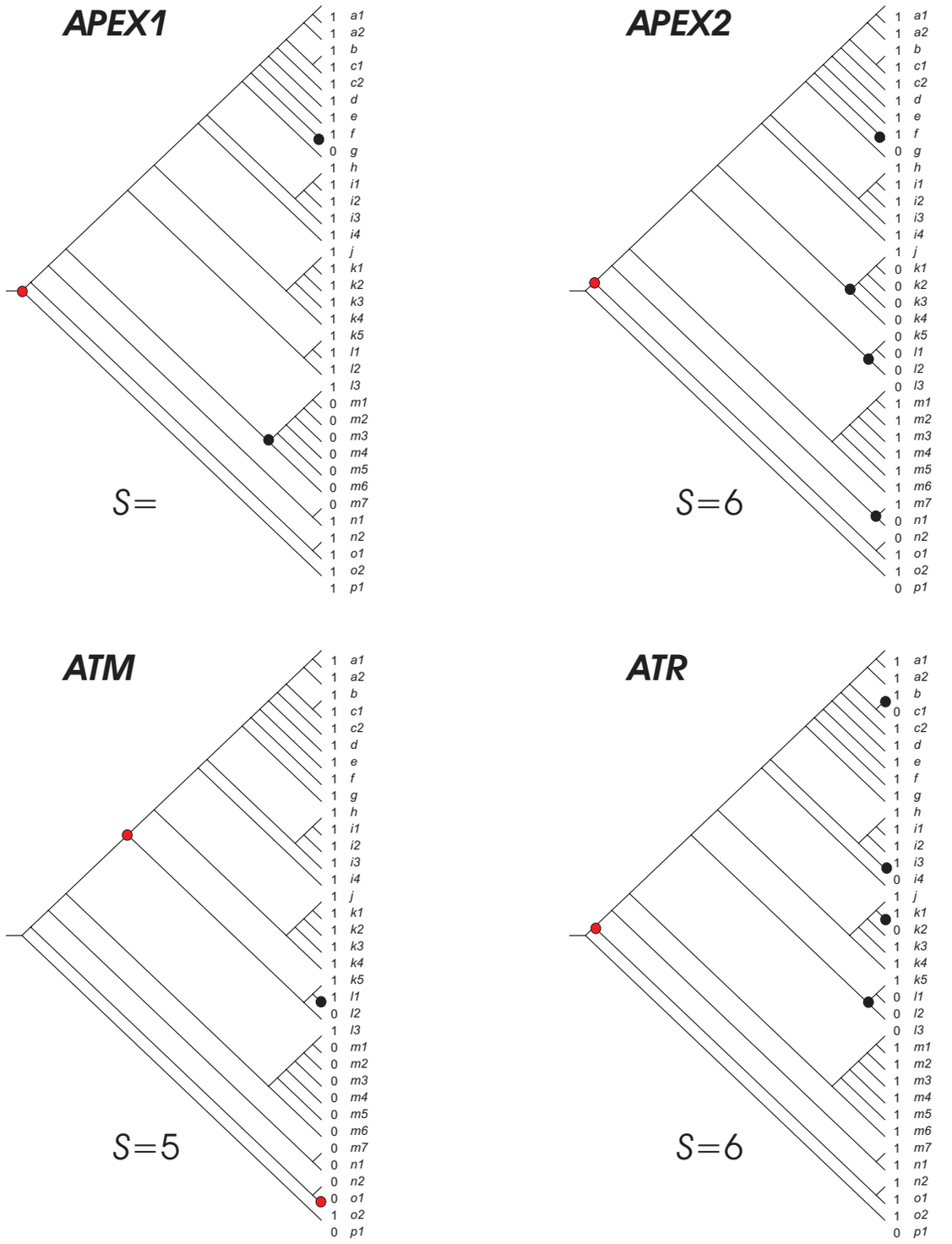
Supplementary Figure S48. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG12441, NOG13097, NOG16176 and NOG21096, as in Supplementary Figure S14.



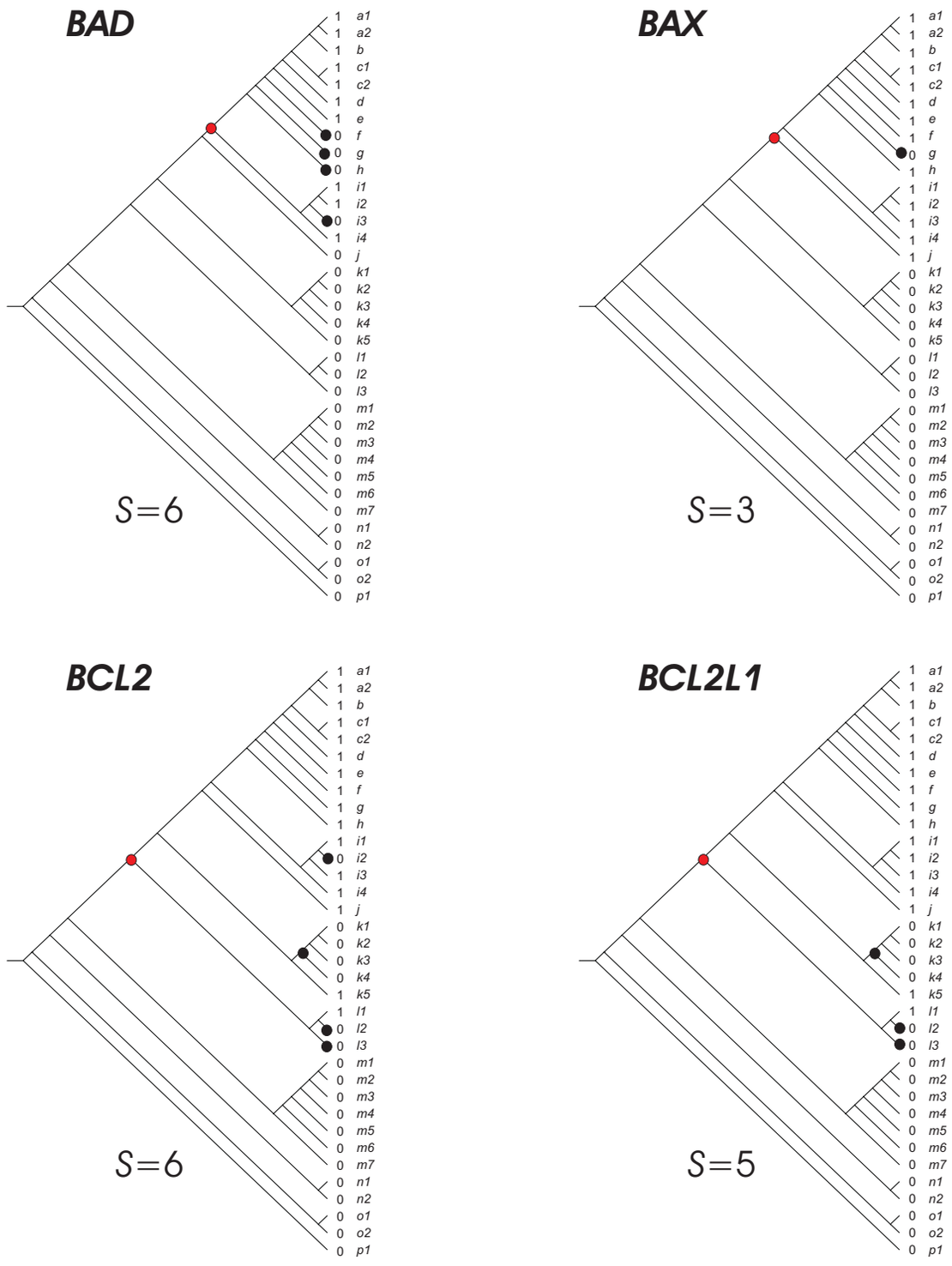
Supplementary Figure S49. Parsimony analysis of eukaryotic clusters of orthologous groups: NOG21098 and NOG36564, as in Supplementary Figure S14.



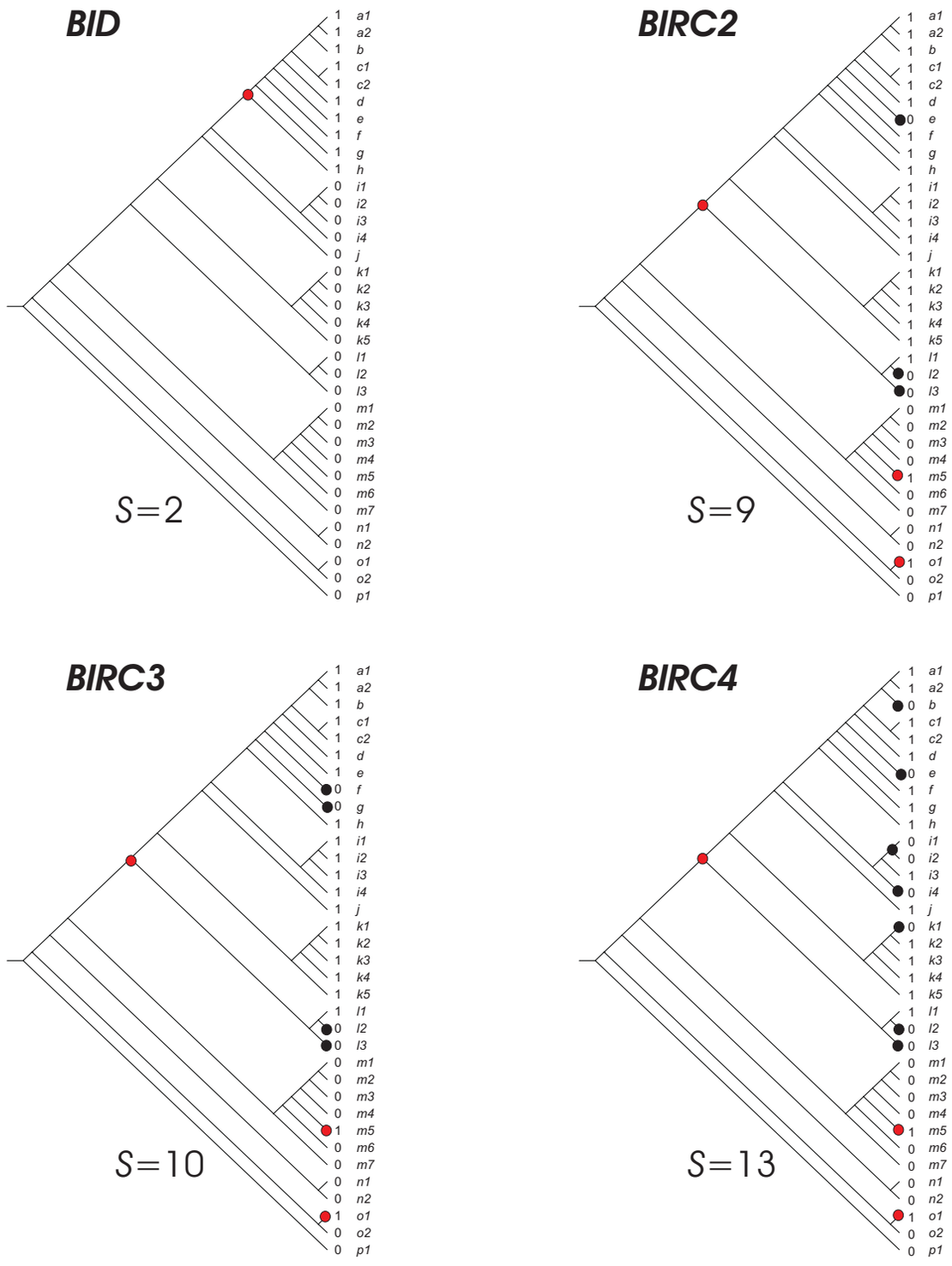
Supplementary Figure S50. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost 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*);



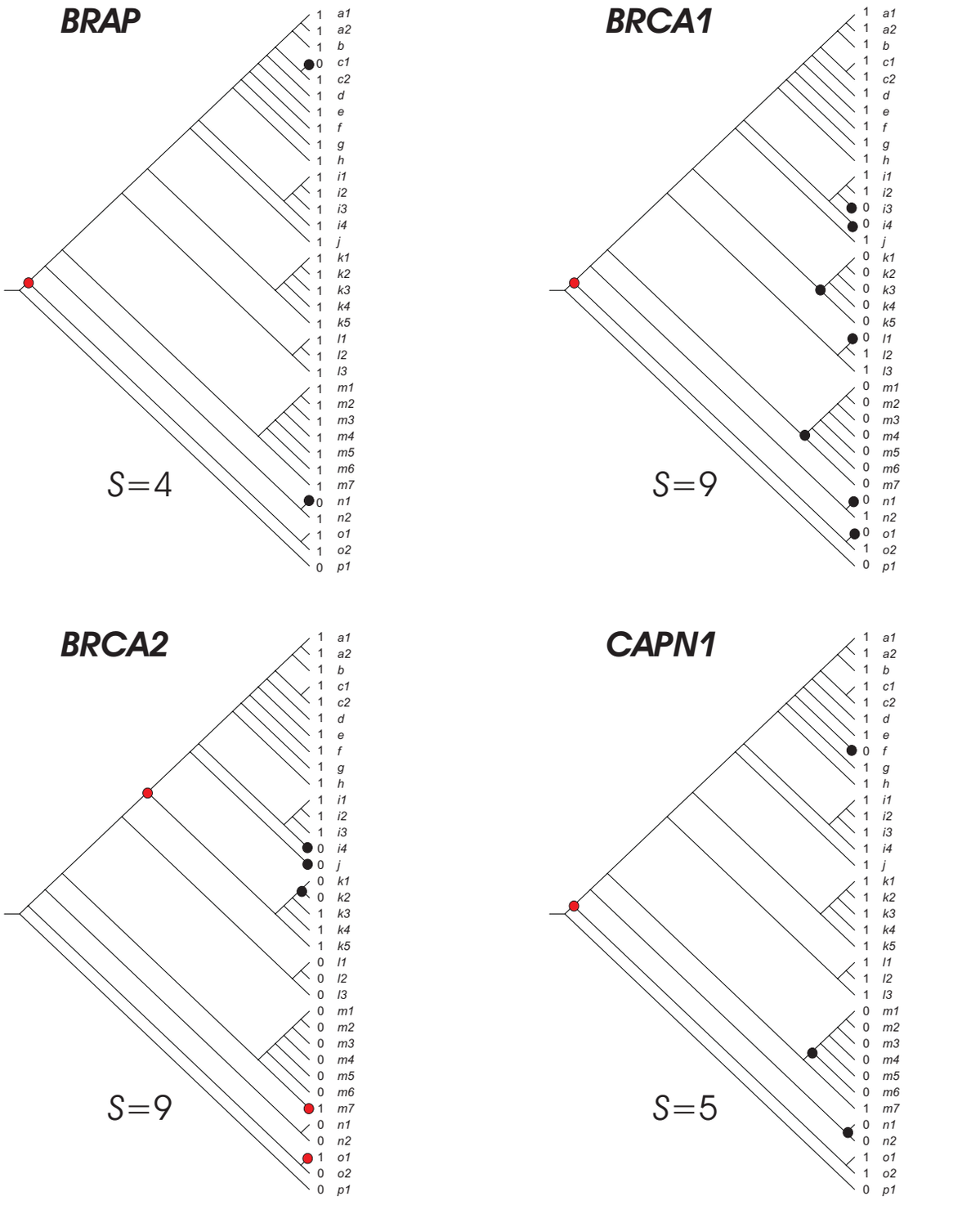
Supplementary Figure S51. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



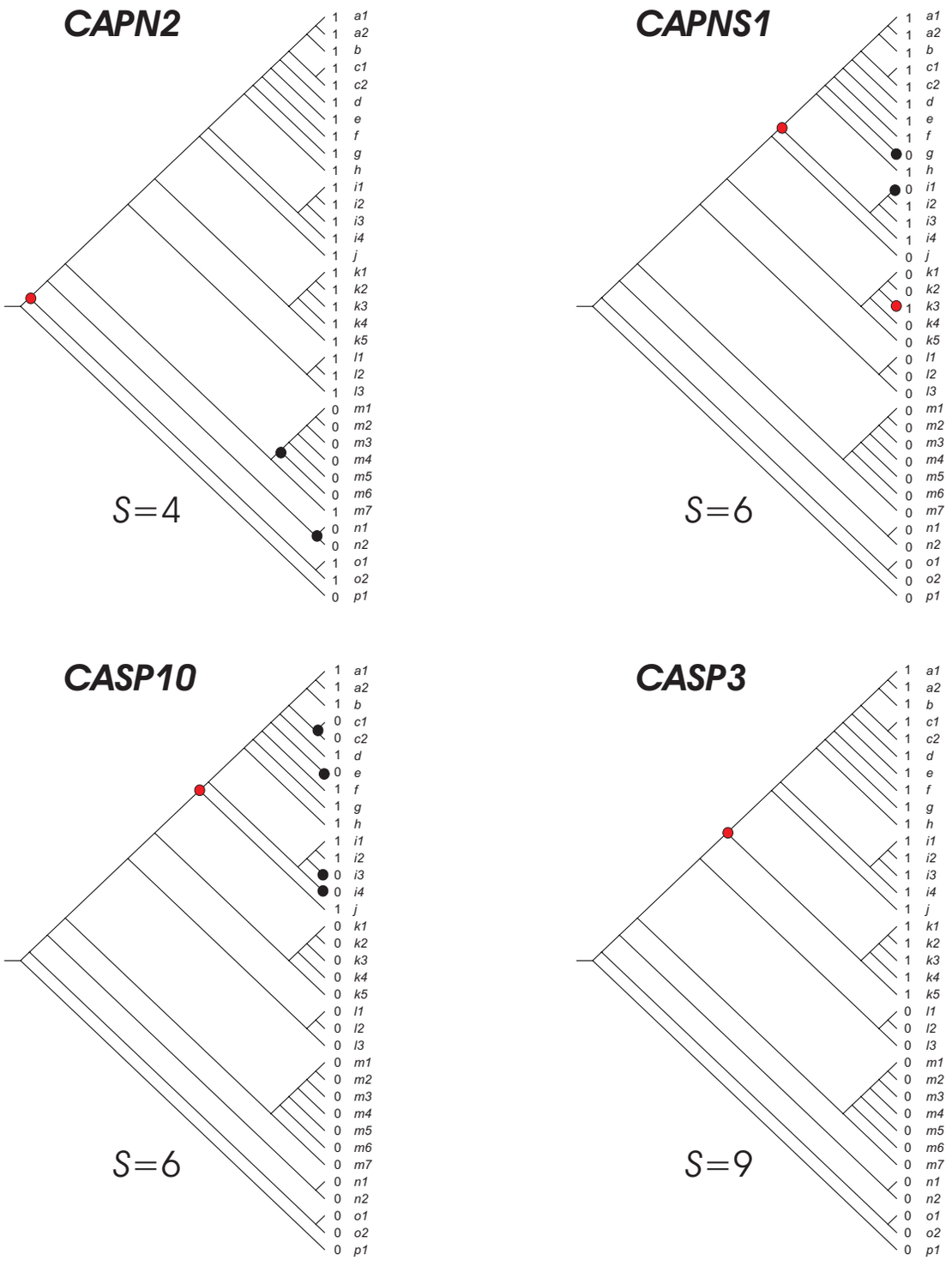
Supplementary Figure S52. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



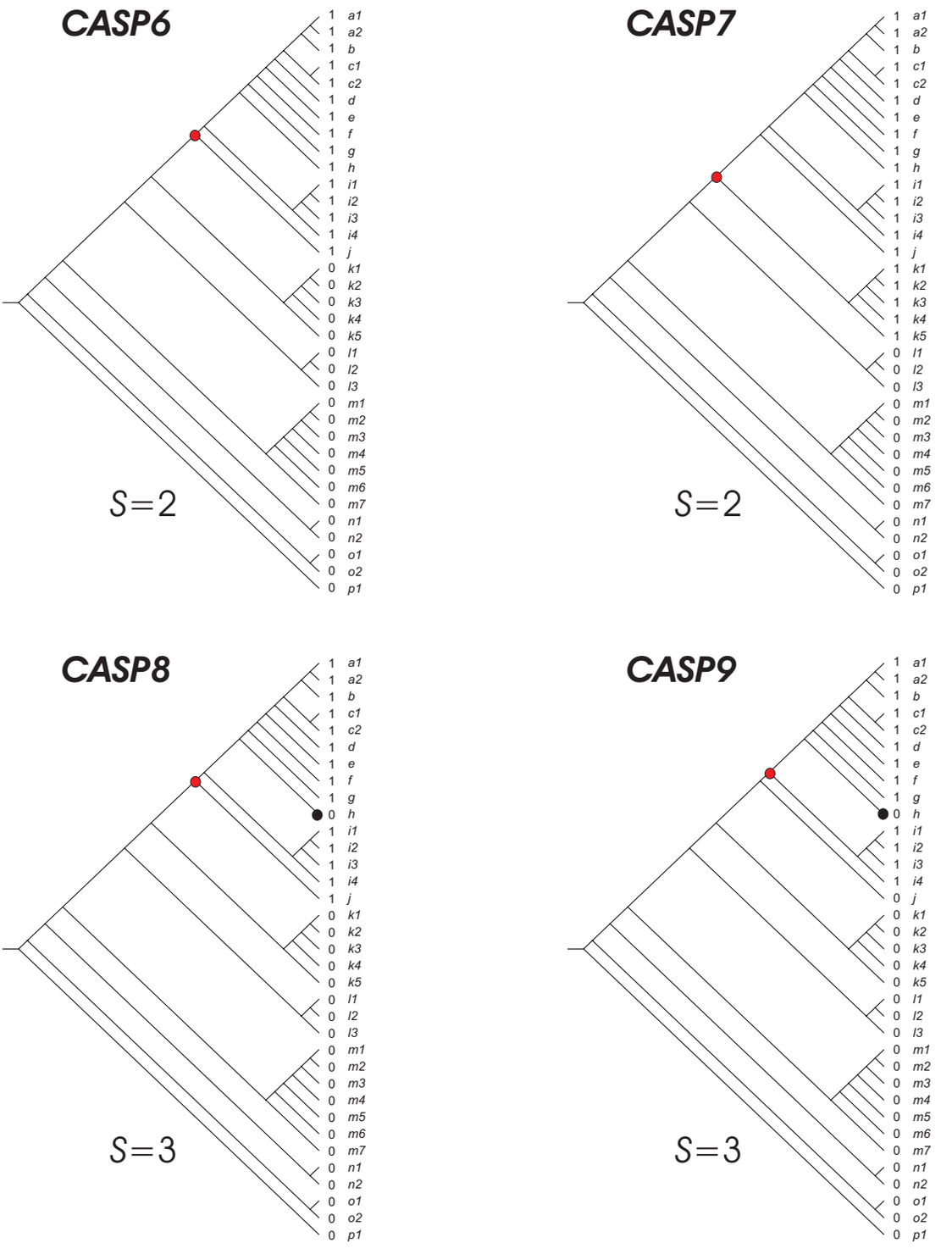
Supplementary Figure S53. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



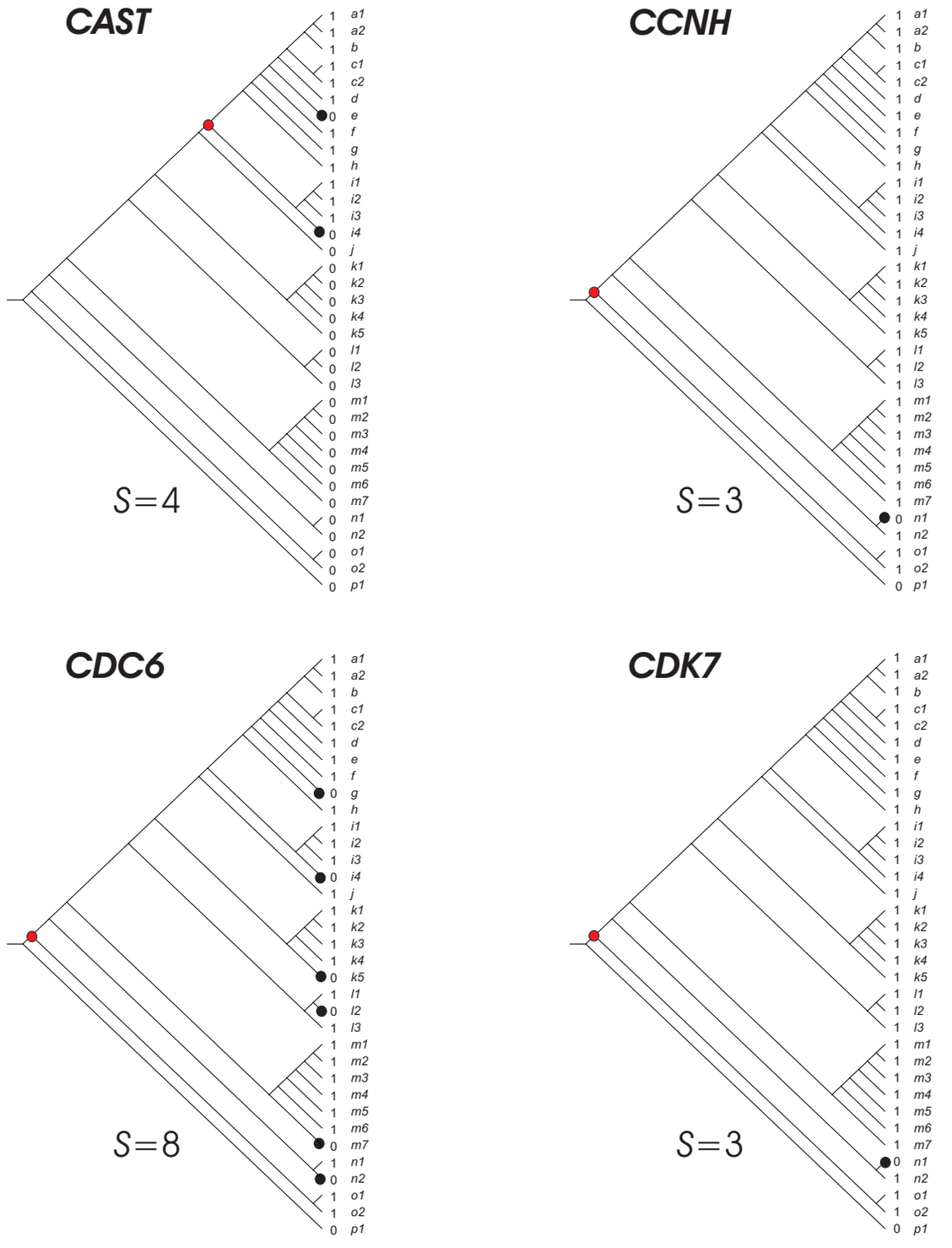
Supplementary Figure S54. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



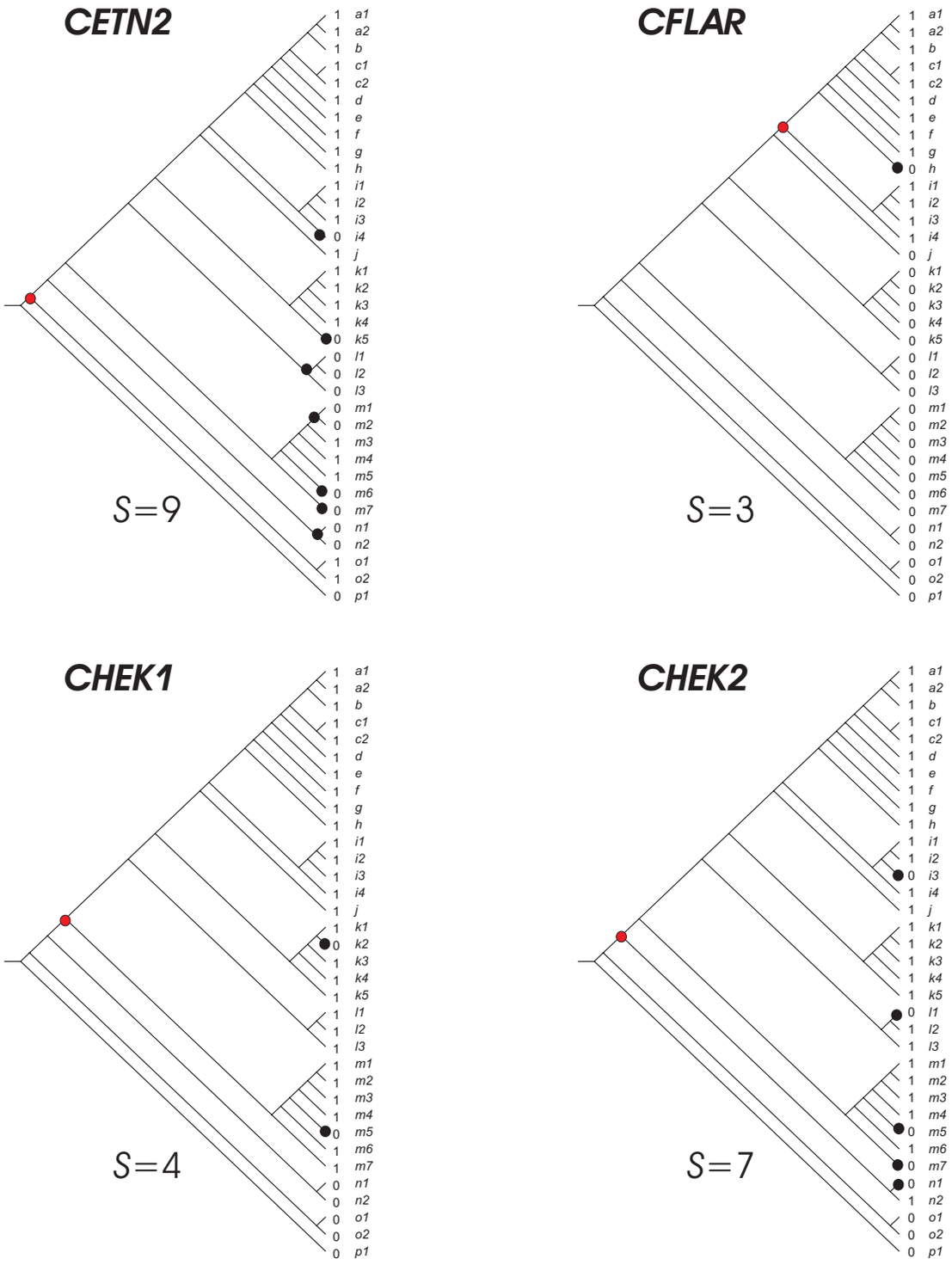
Supplementary Figure S55. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



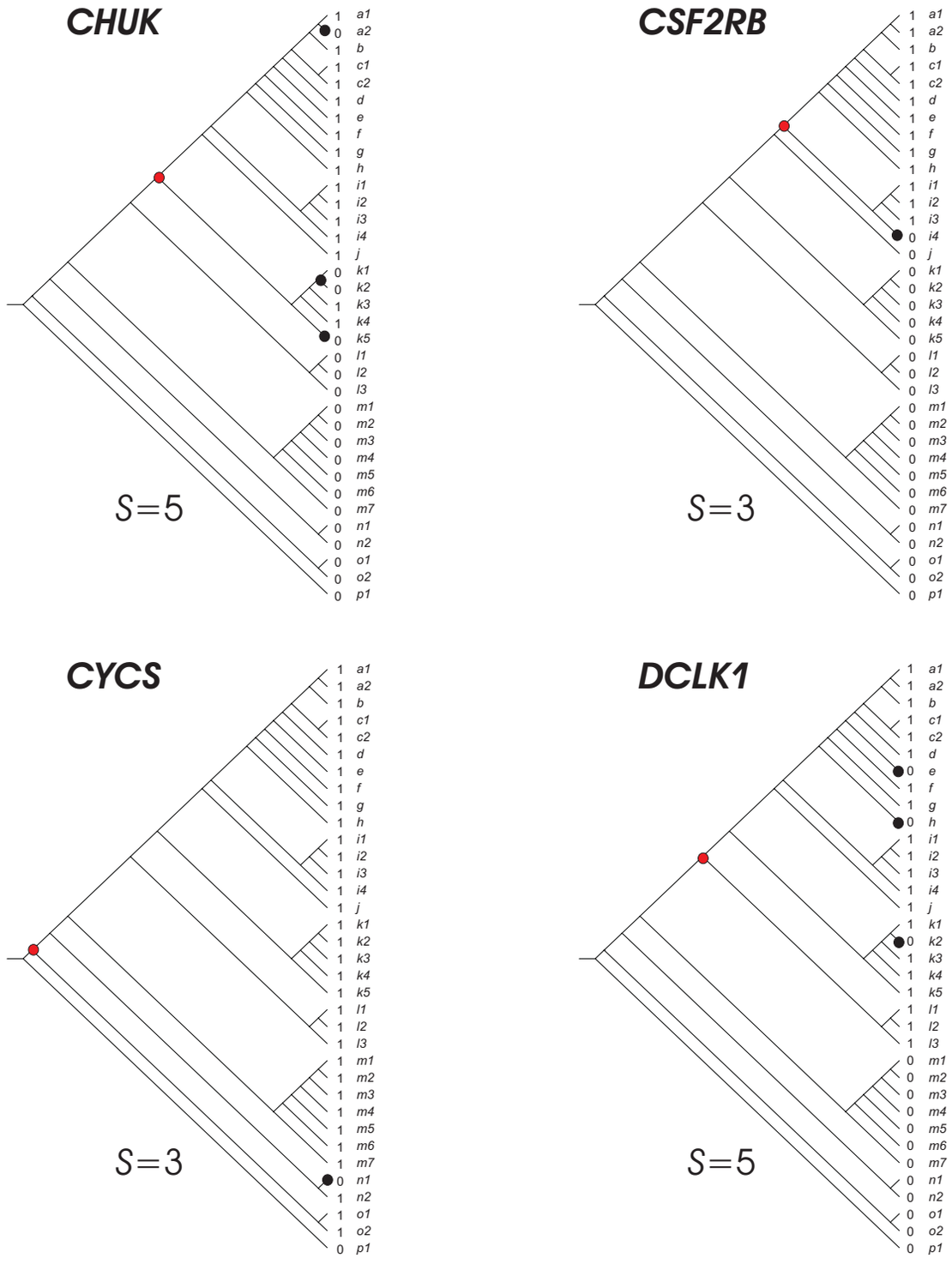
Supplementary Figure S56. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



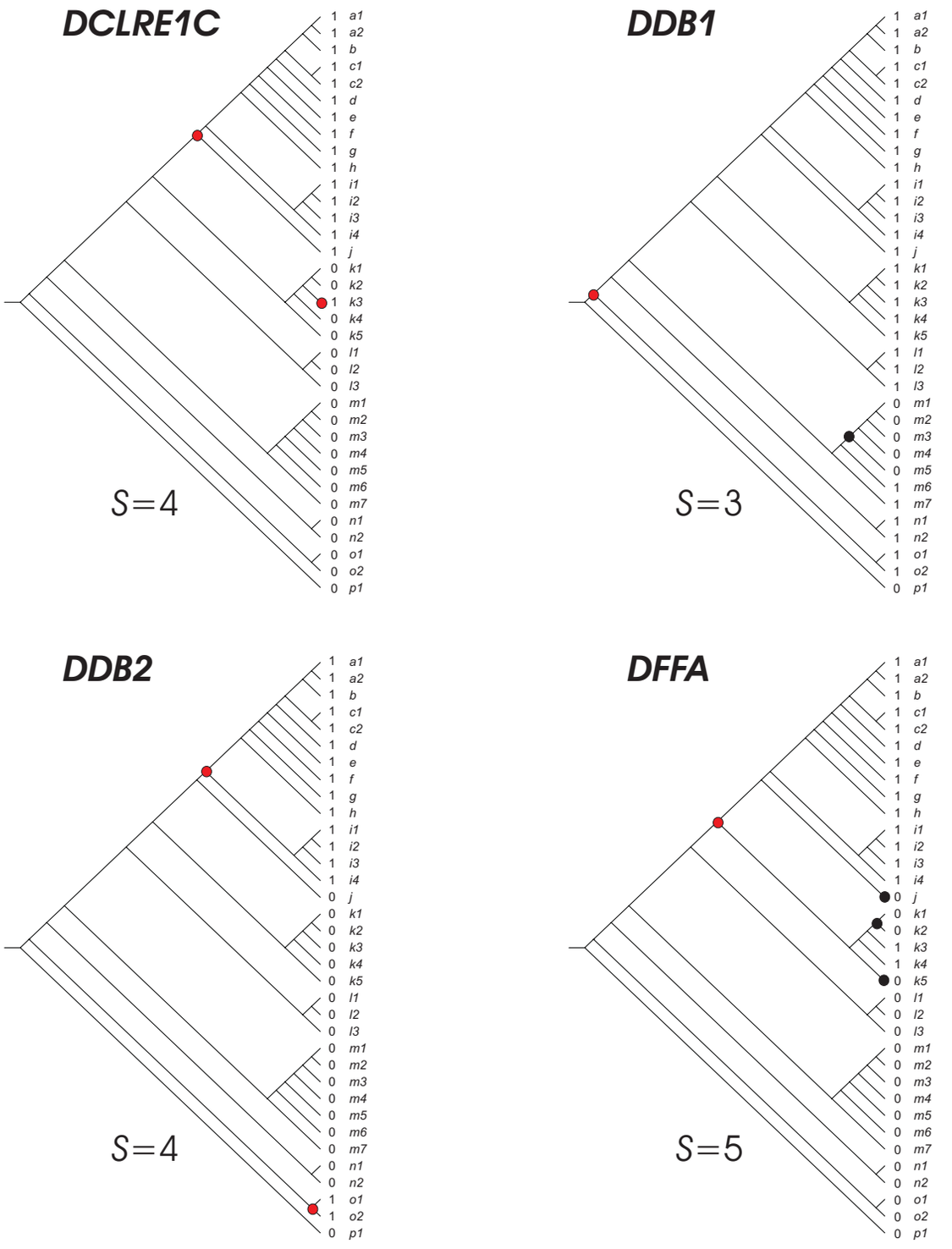
Supplementary Figure S57. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



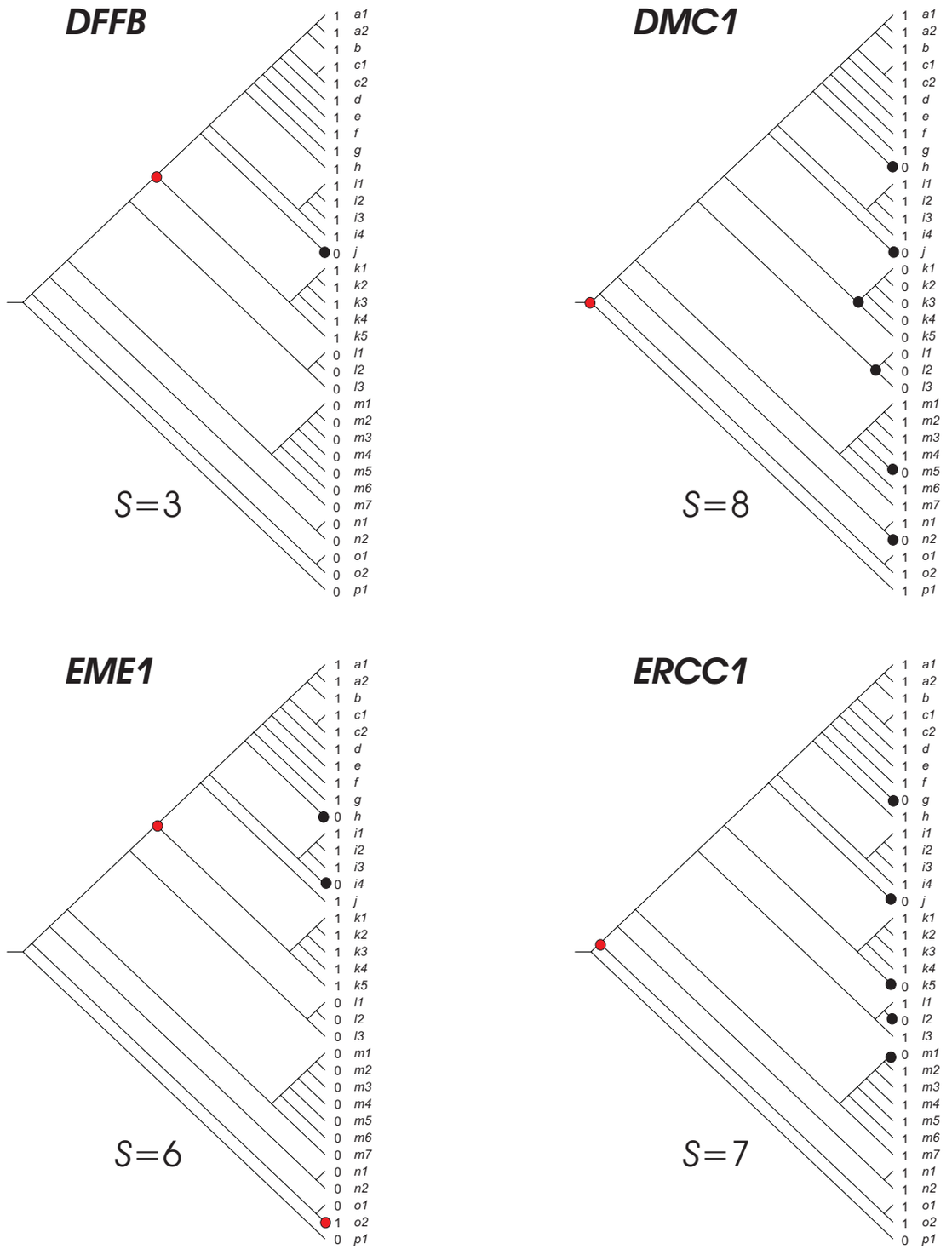
Supplementary Figure S58. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



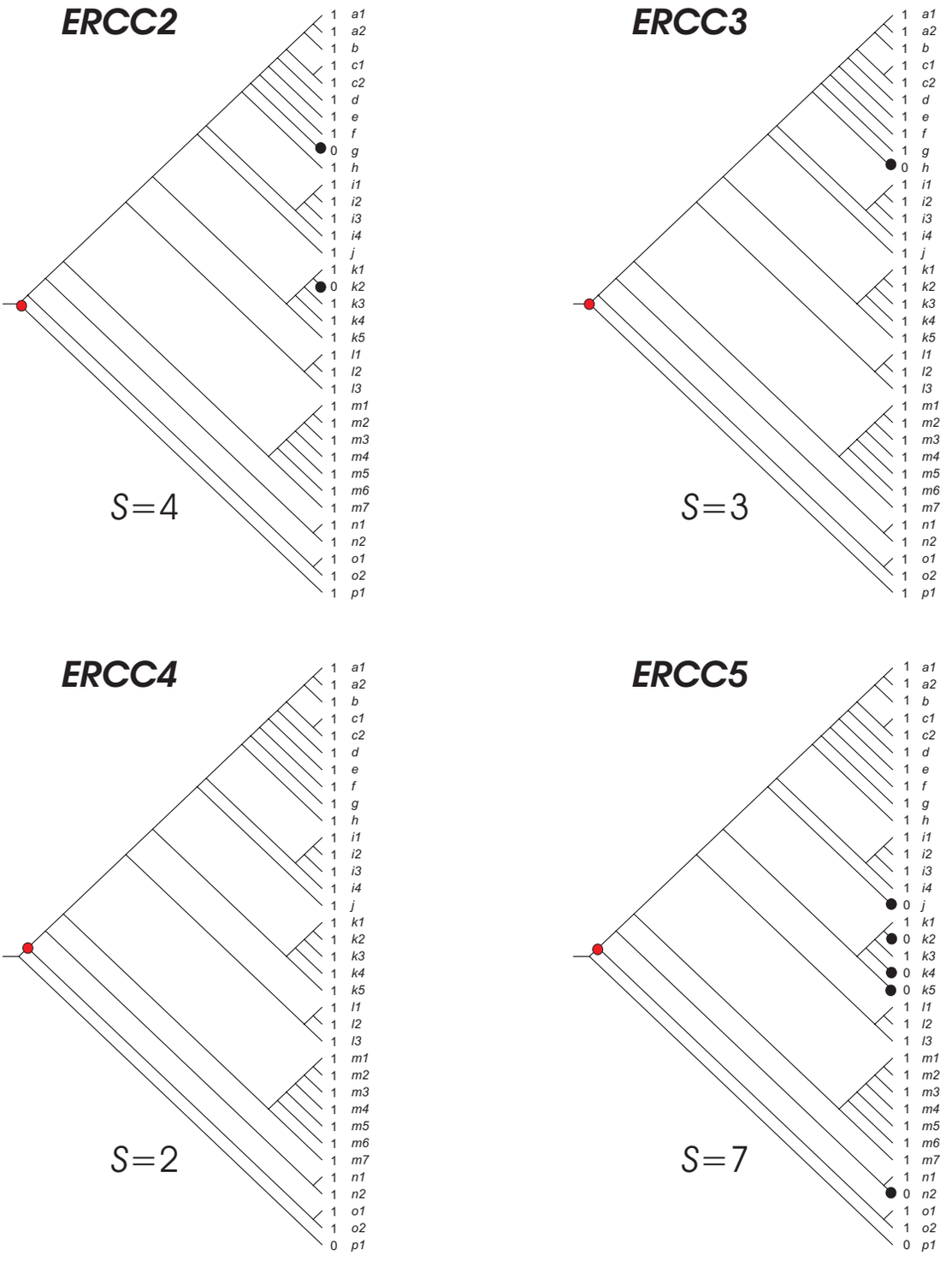
Supplementary Figure S59. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



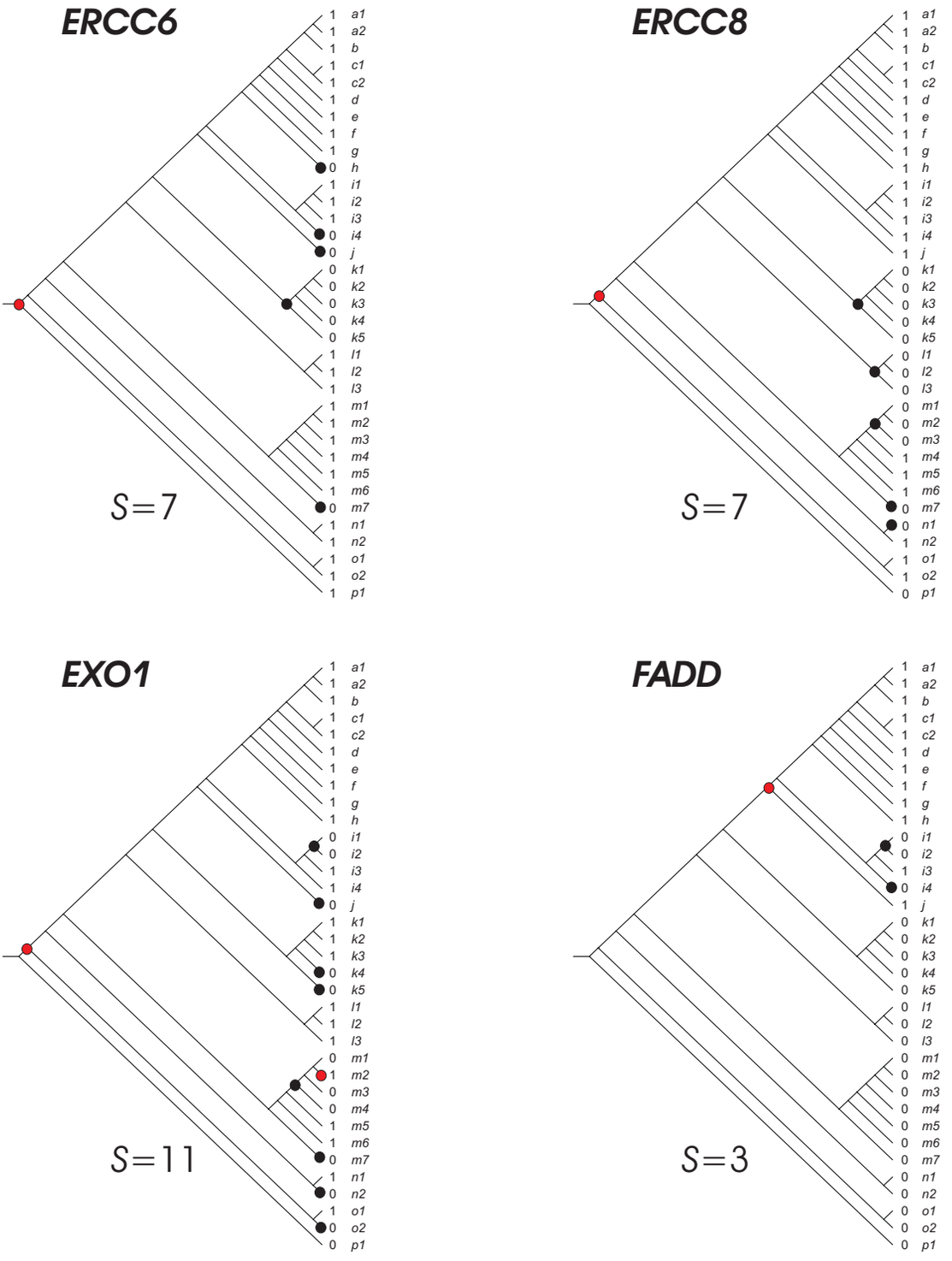
Supplementary Figure S60. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



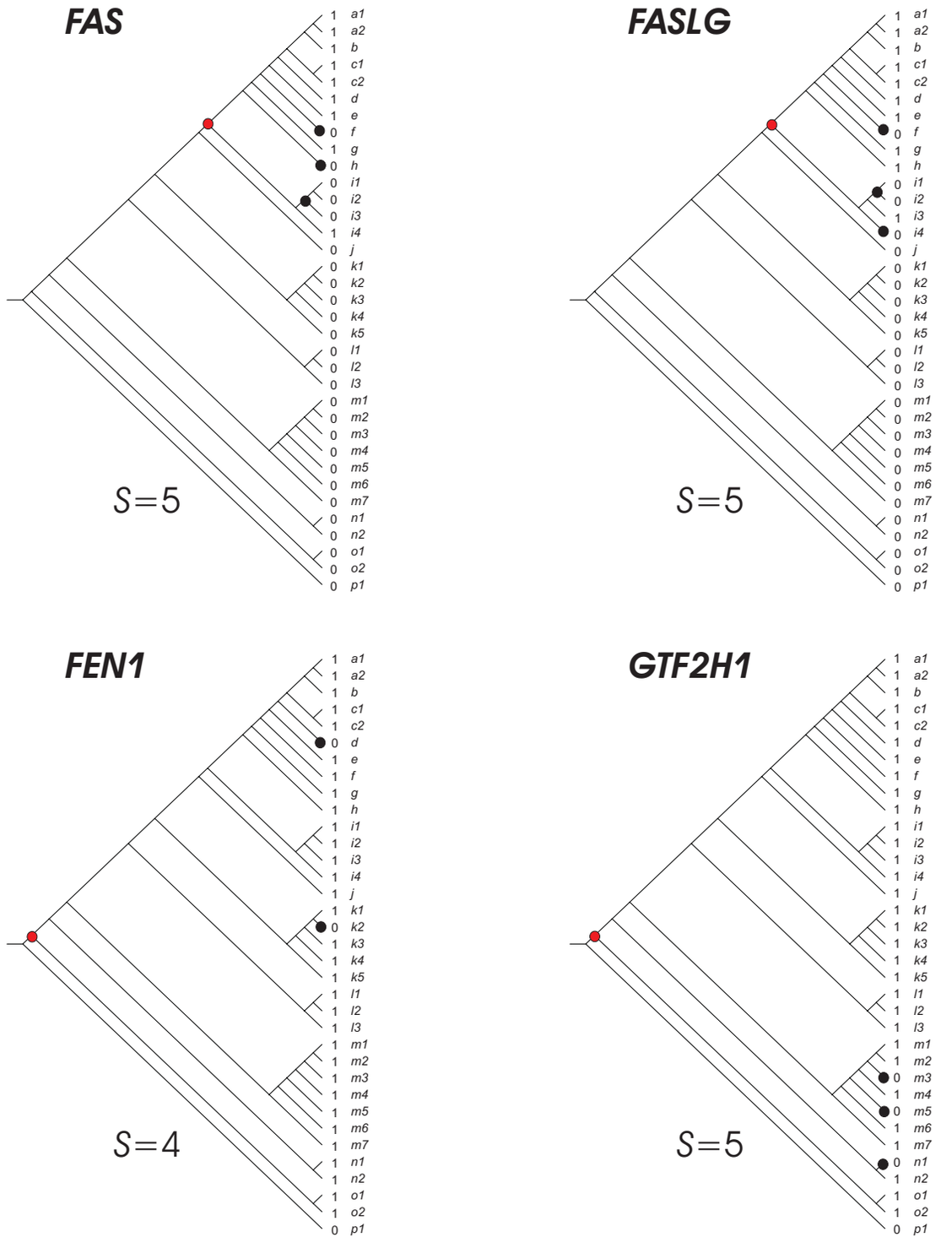
Supplementary Figure S61. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



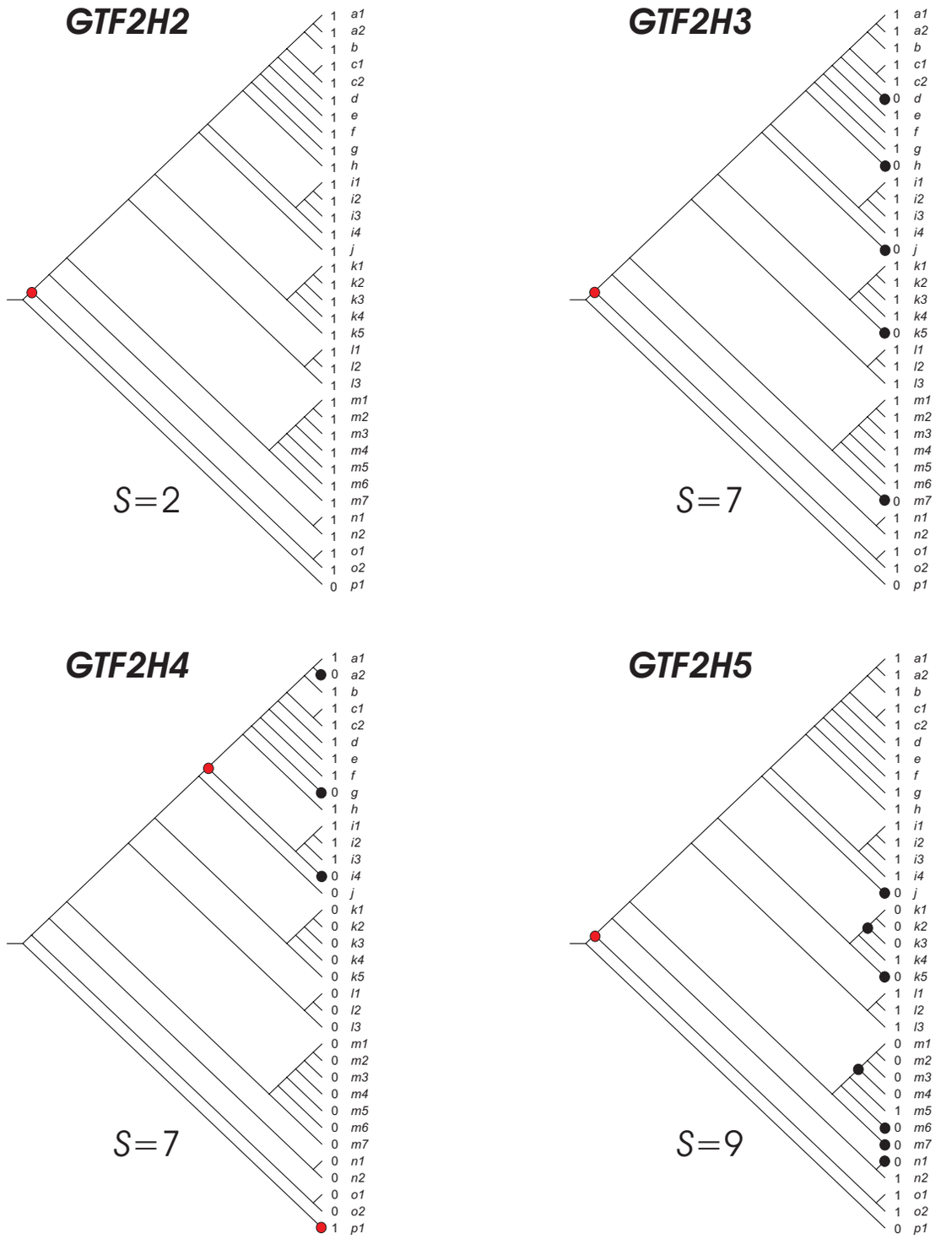
Supplementary Figure S62. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



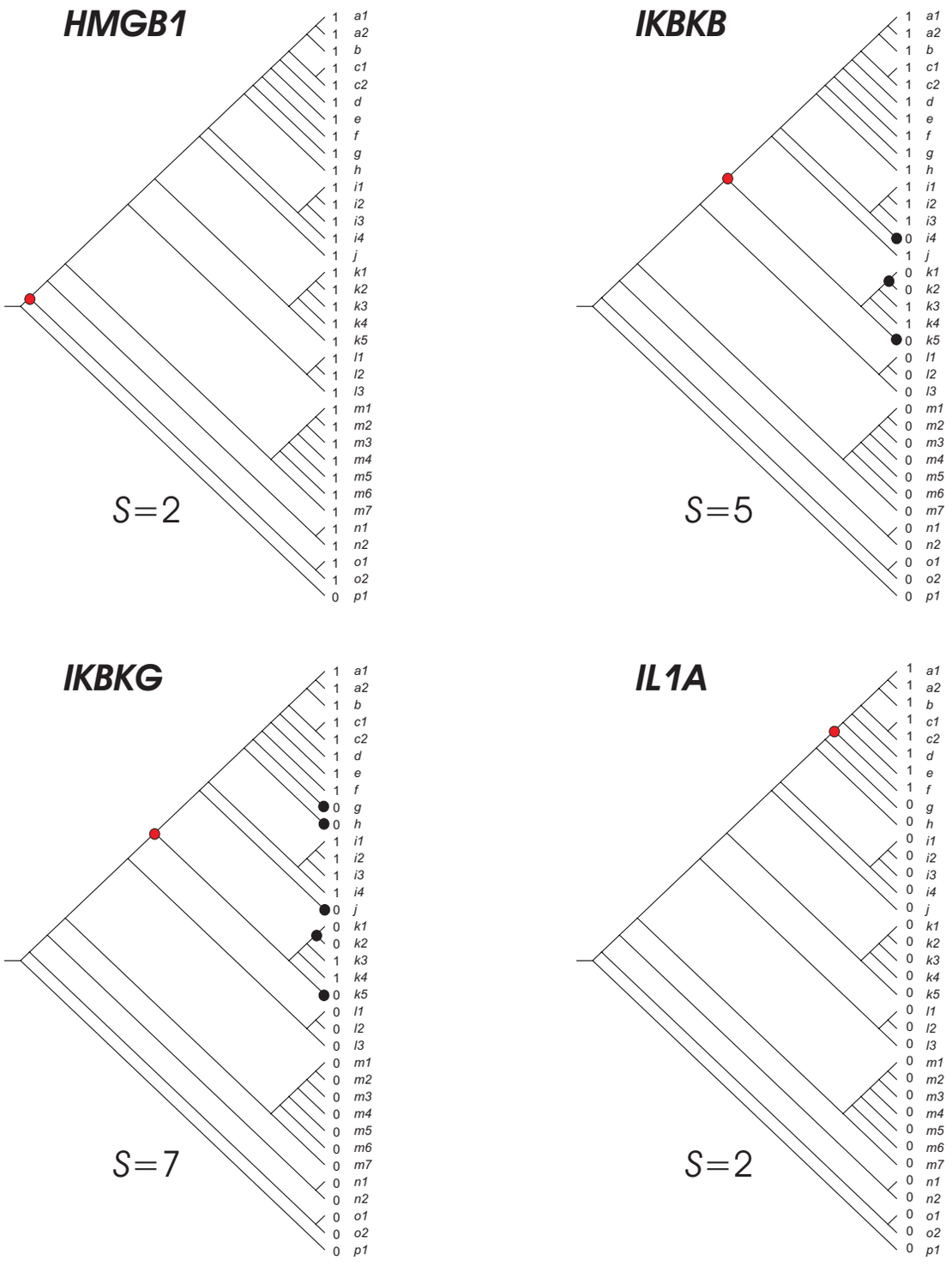
Supplementary Figure S63. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



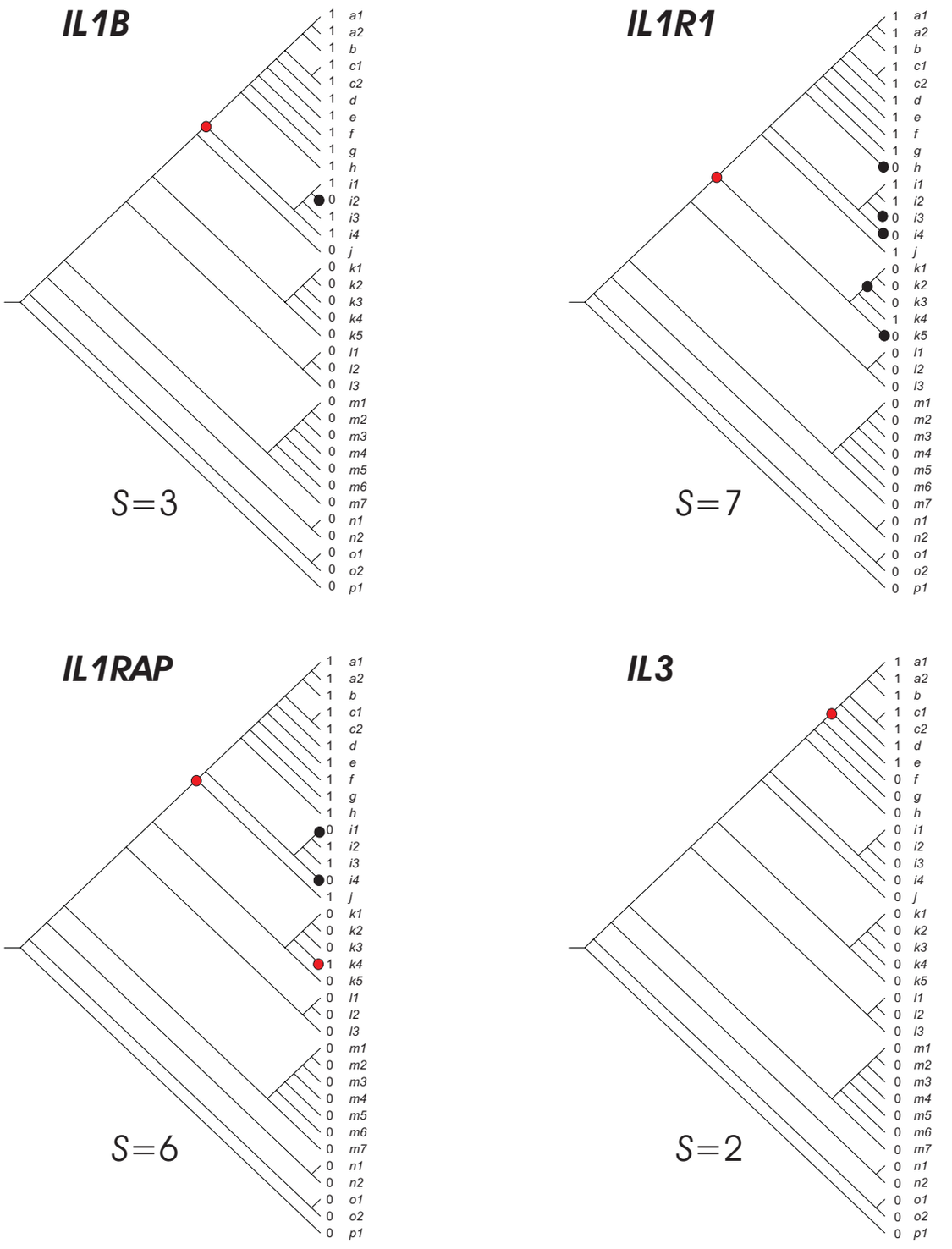
Supplementary Figure S64. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



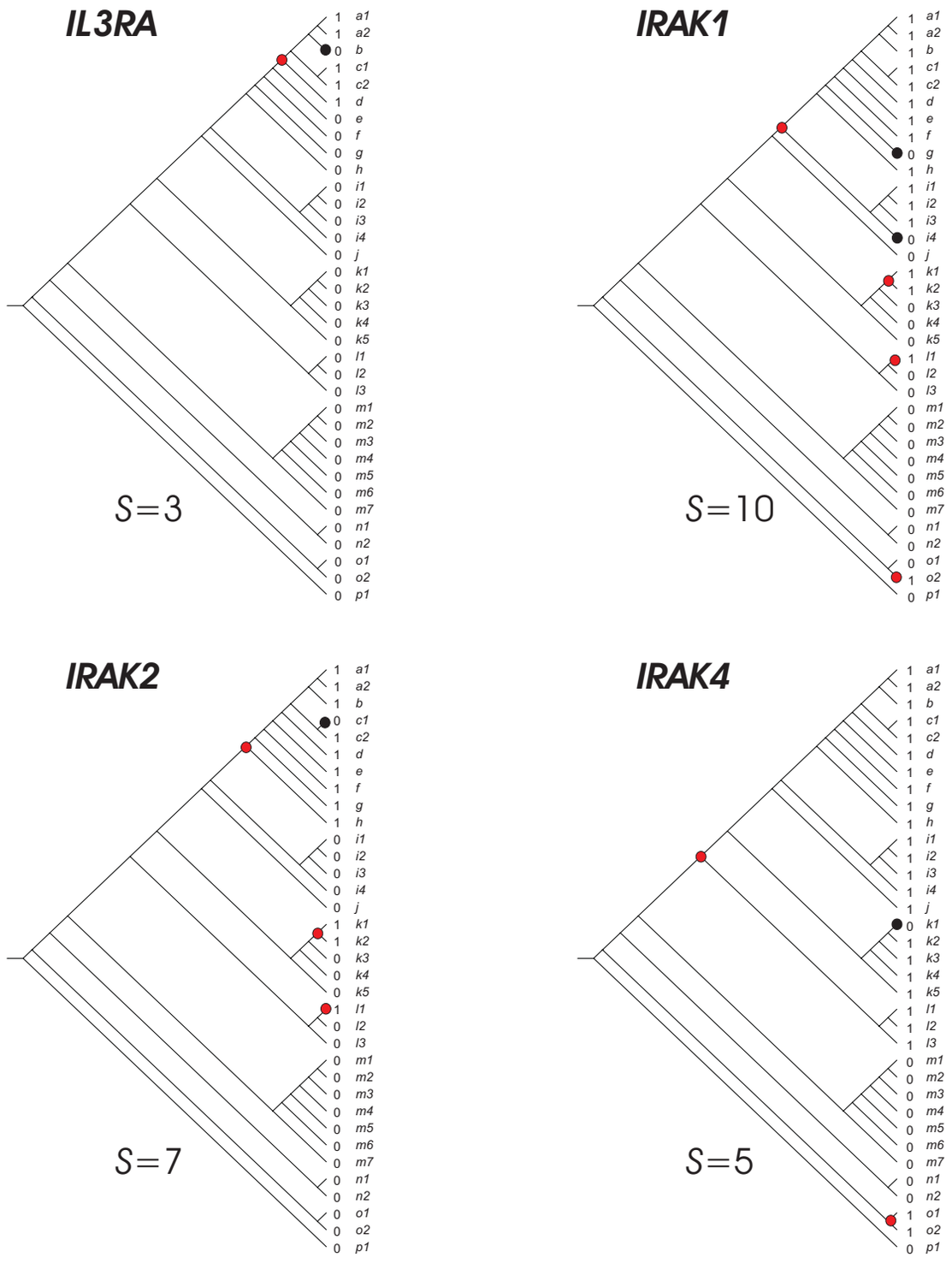
Supplementary Figure S65. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



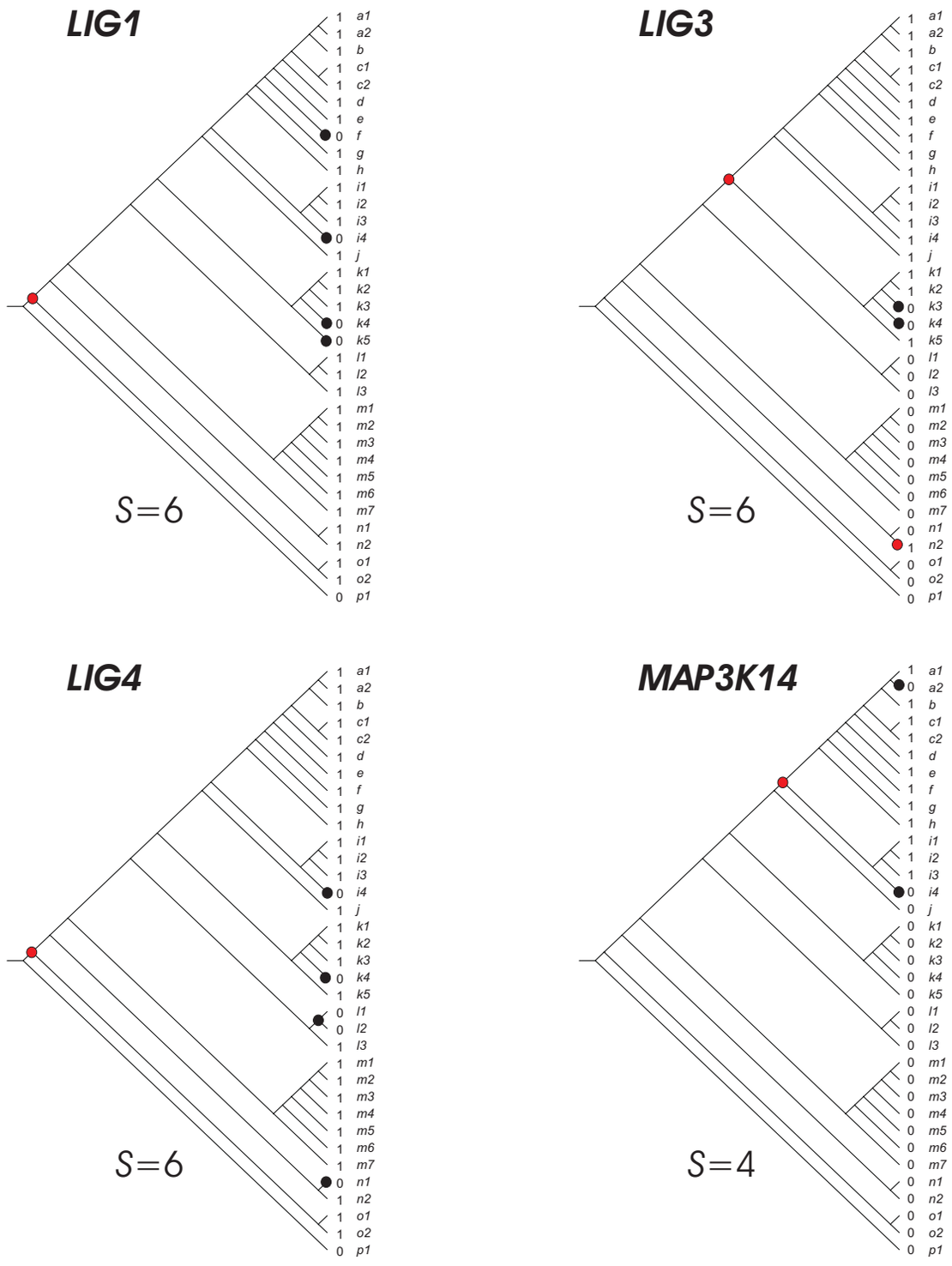
Supplementary Figure S66. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



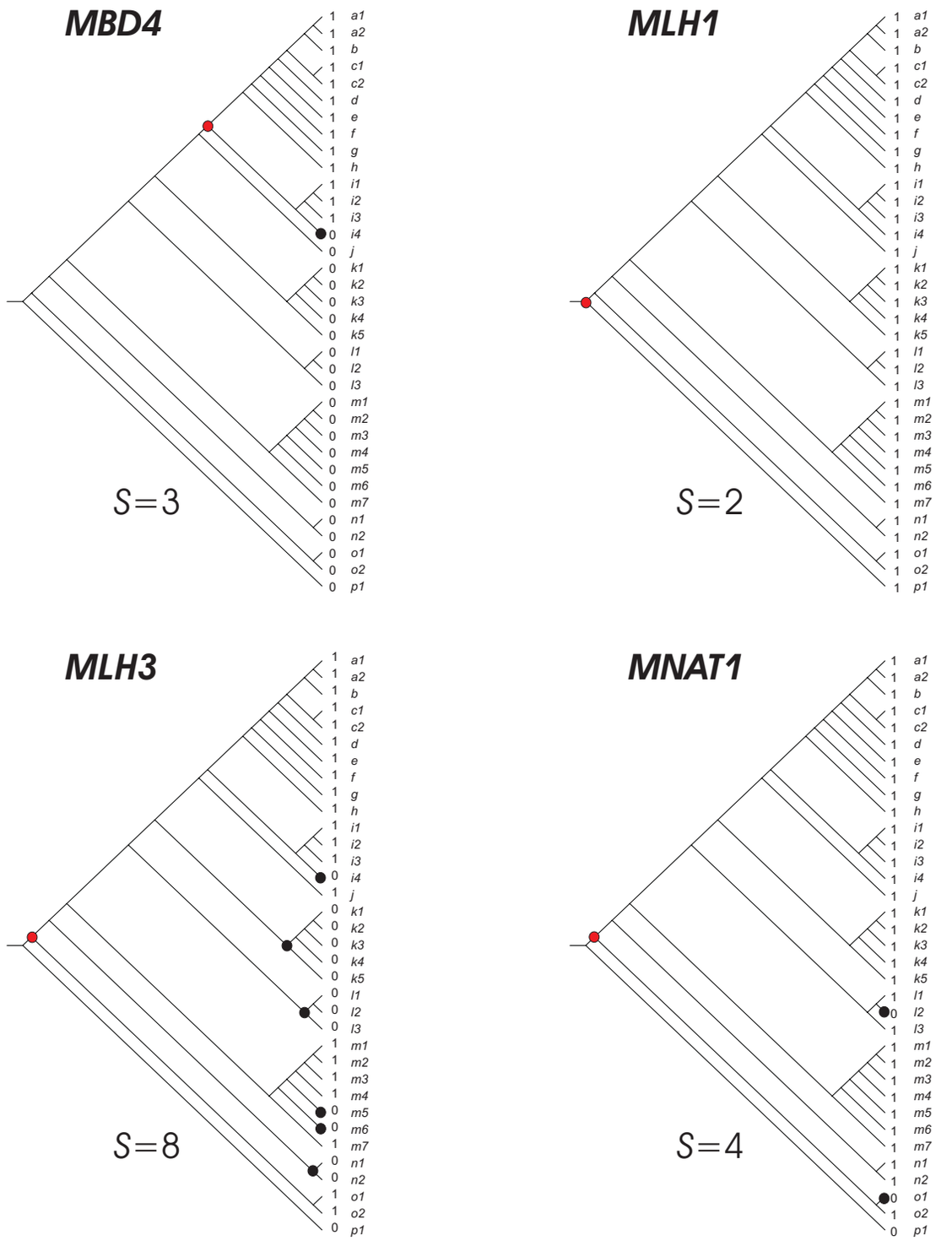
Supplementary Figure S67. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



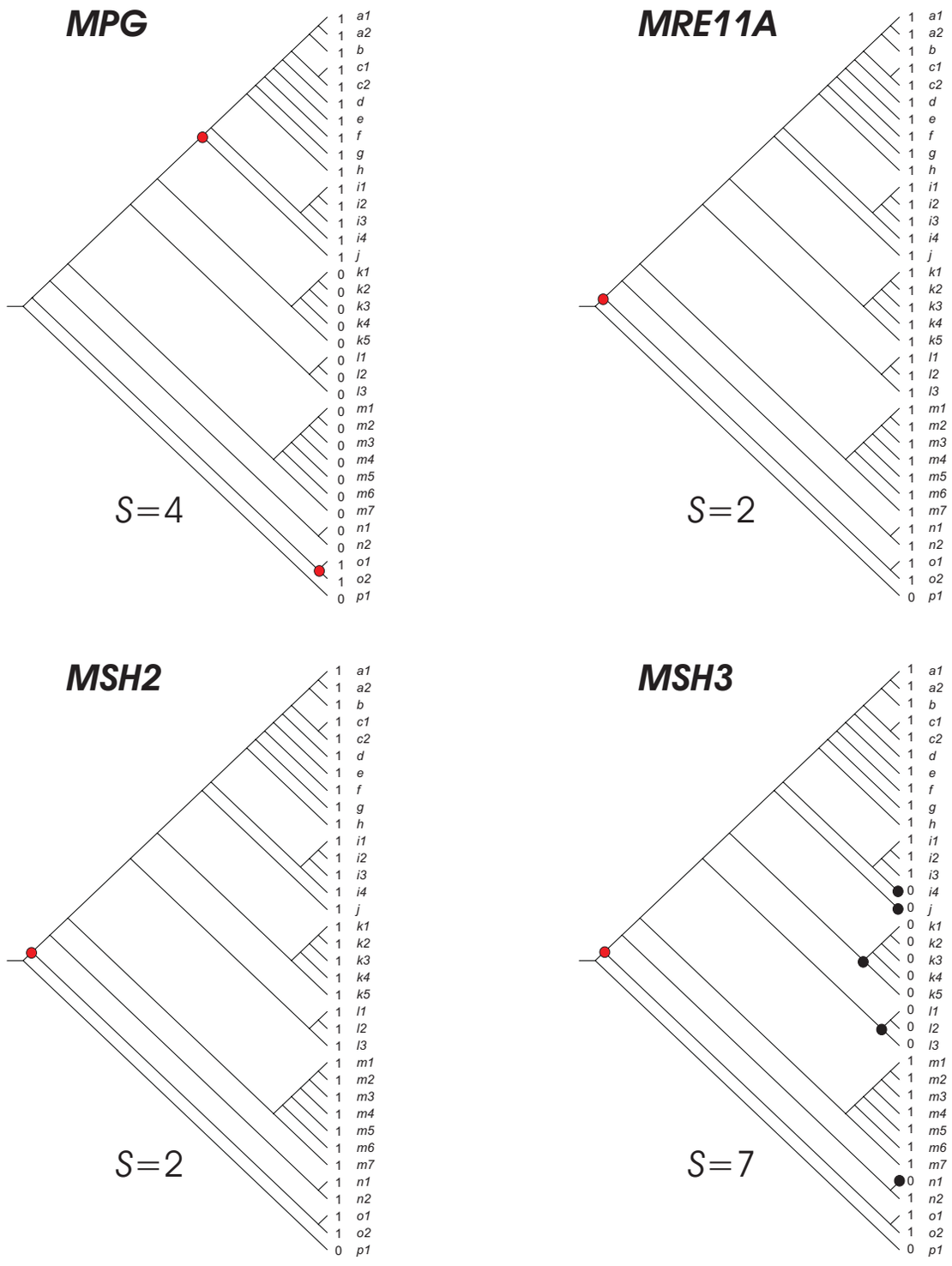
Supplementary Figure S68. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



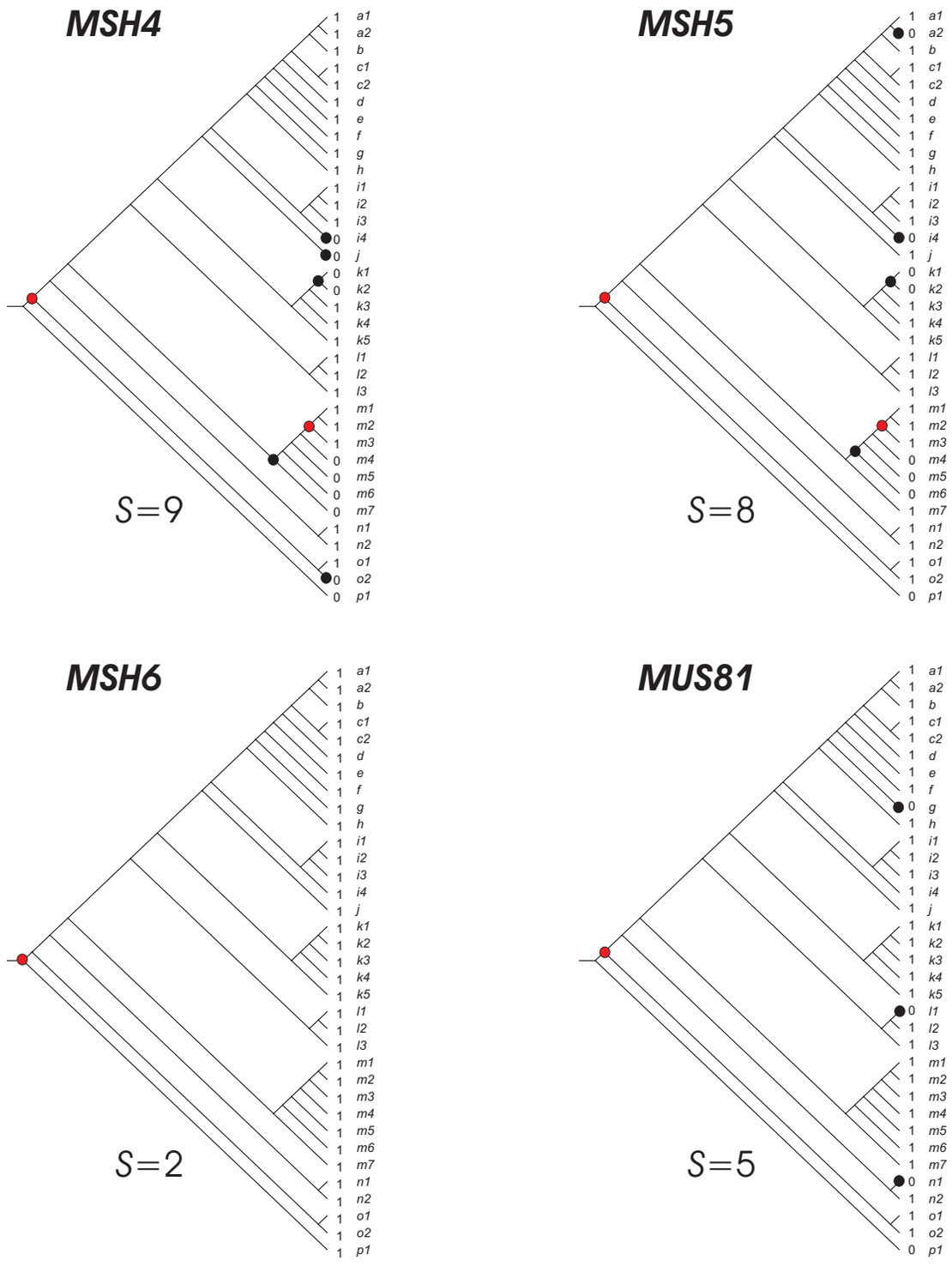
Supplementary Figure S69. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



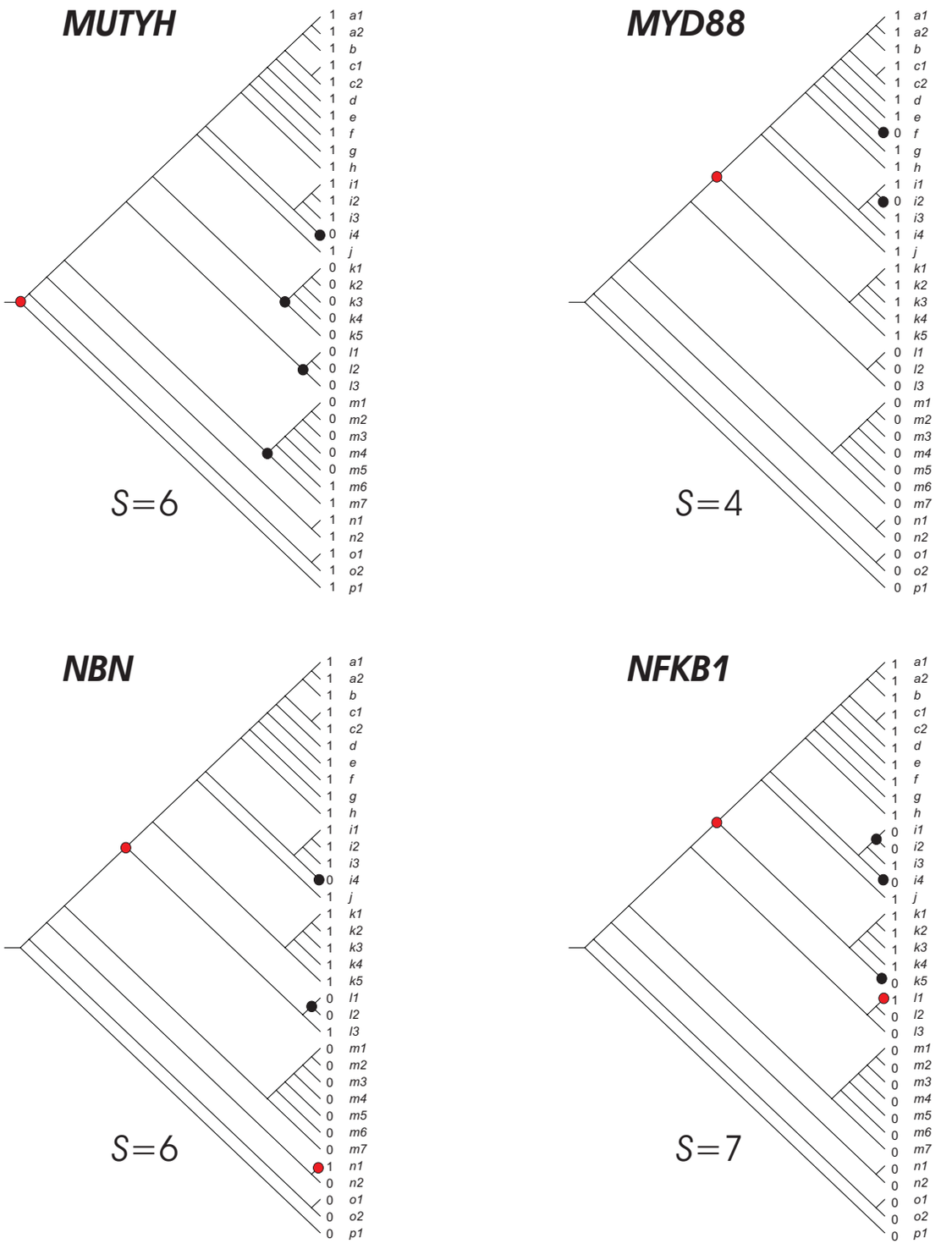
Supplementary Figure S70. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



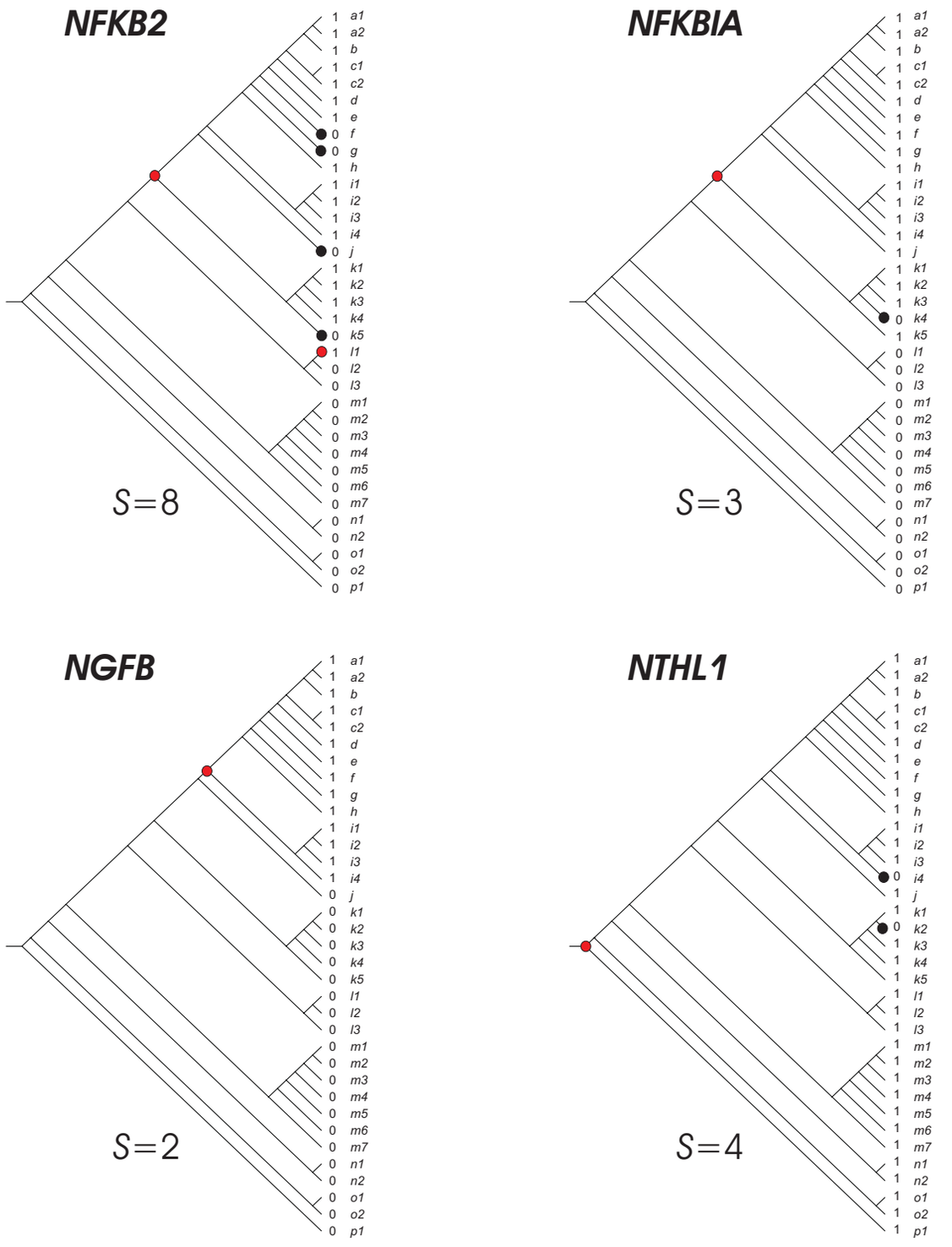
Supplementary Figure S71. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



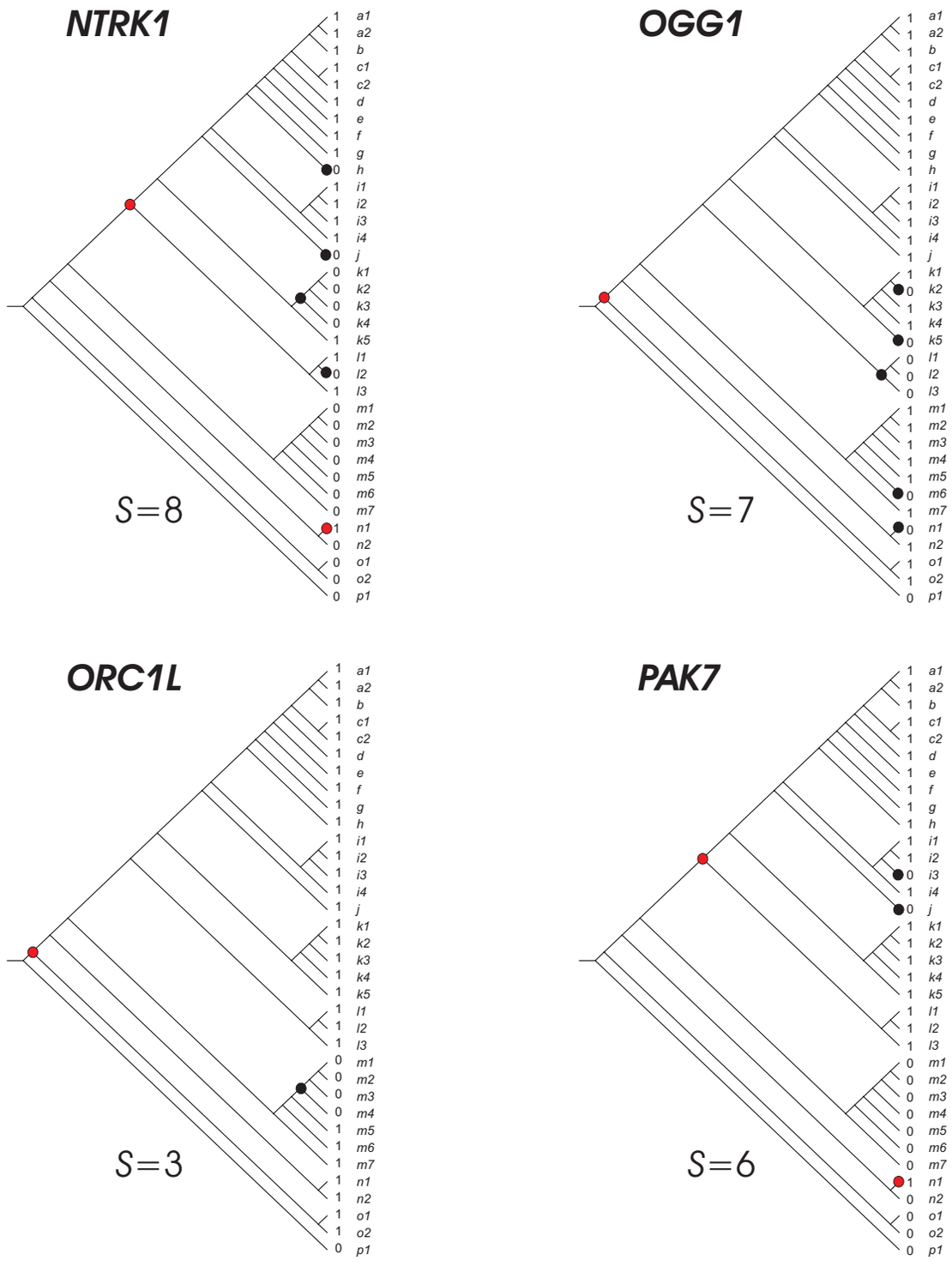
Supplementary Figure S72. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



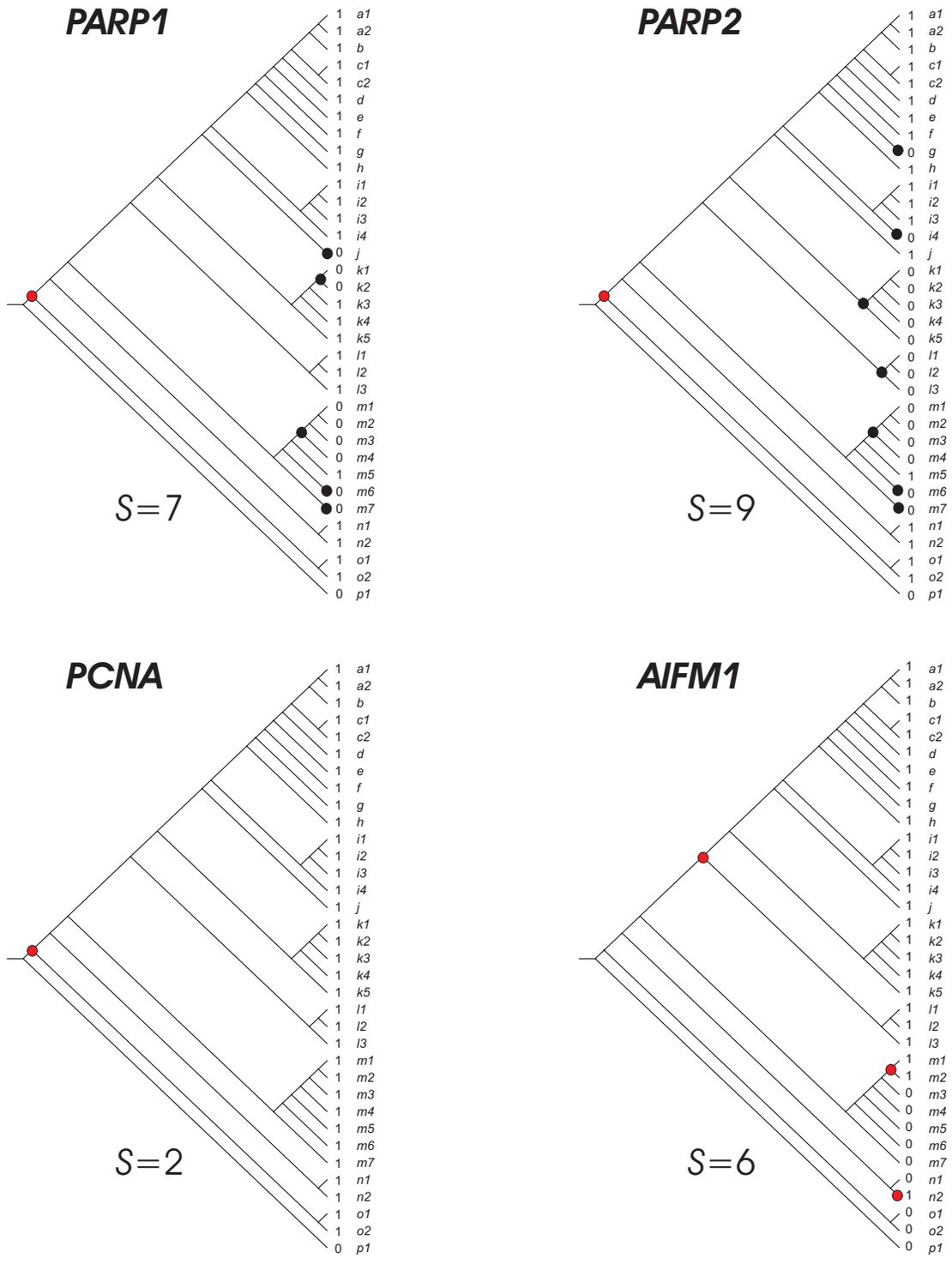
Supplementary Figure S73. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



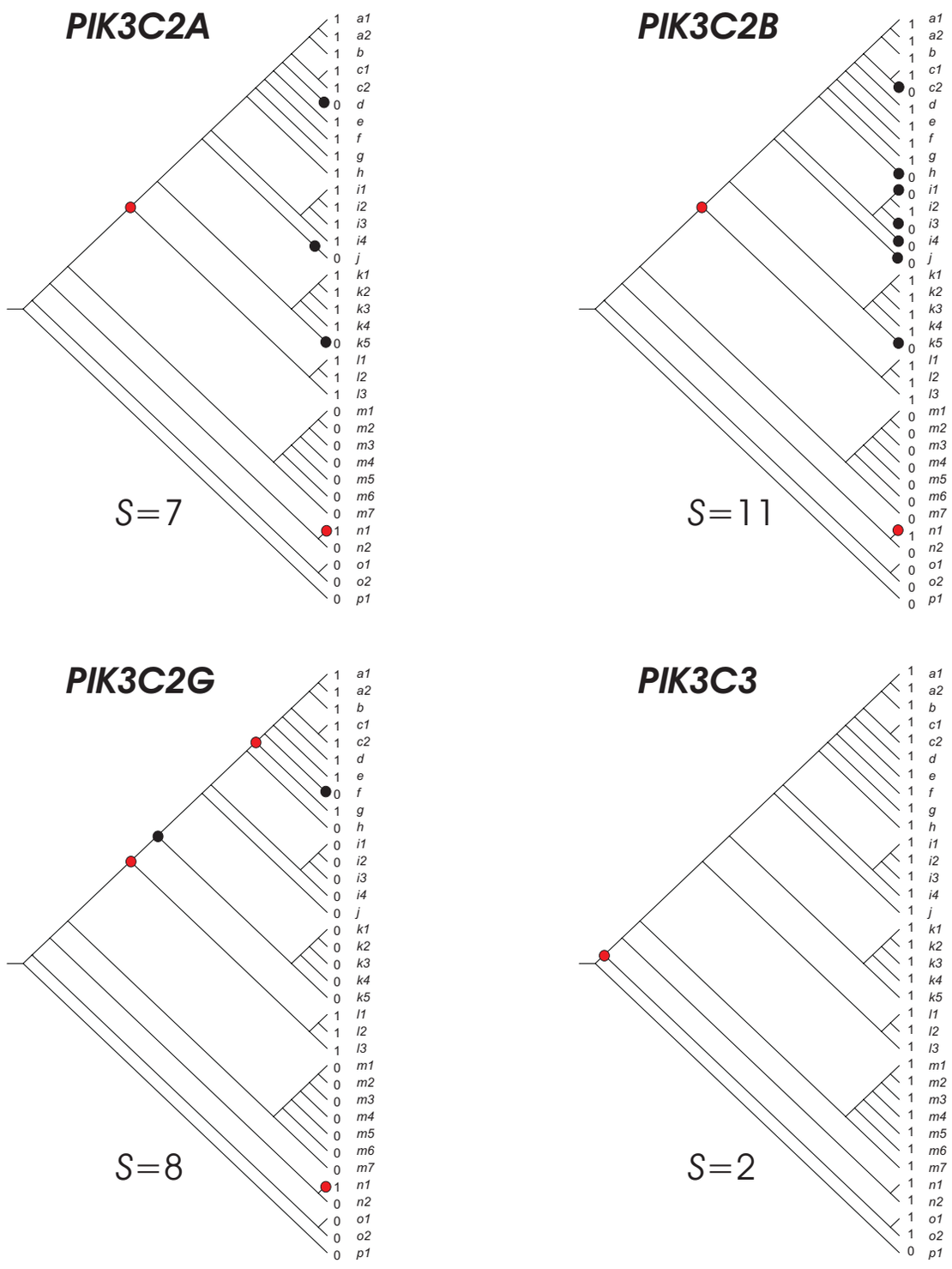
Supplementary Figure S74. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



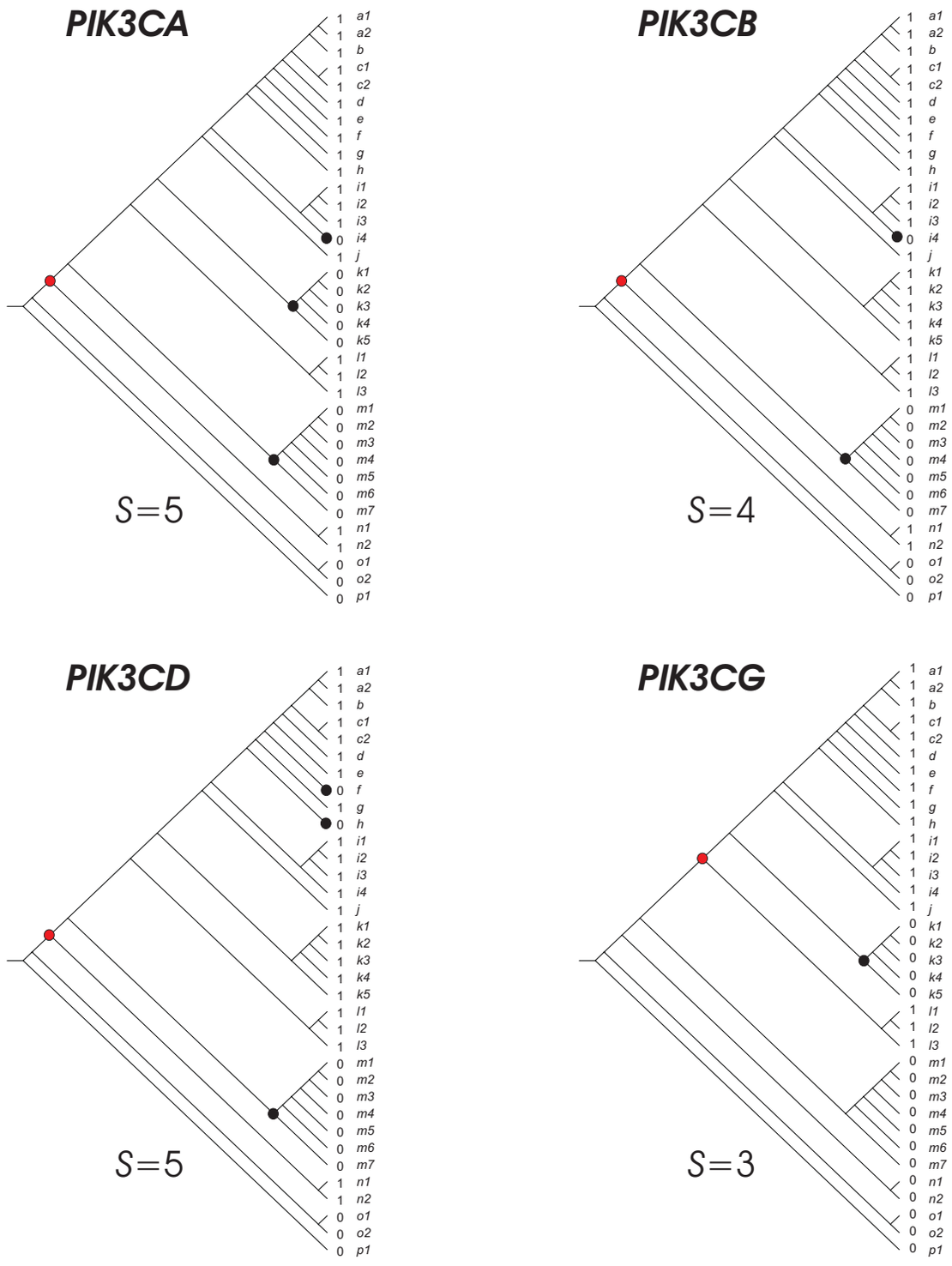
Supplementary Figure S75. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



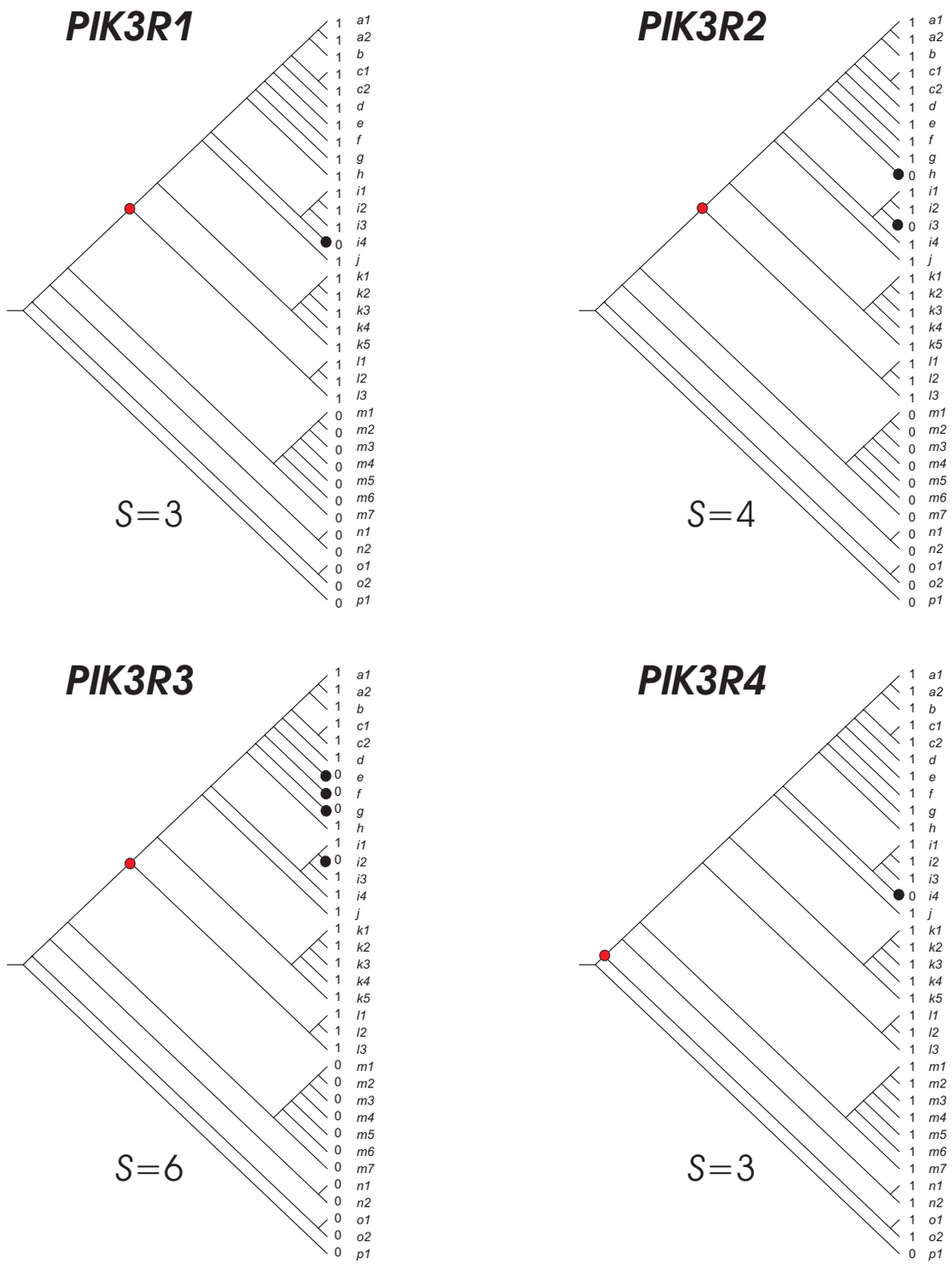
Supplementary Figure S76. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



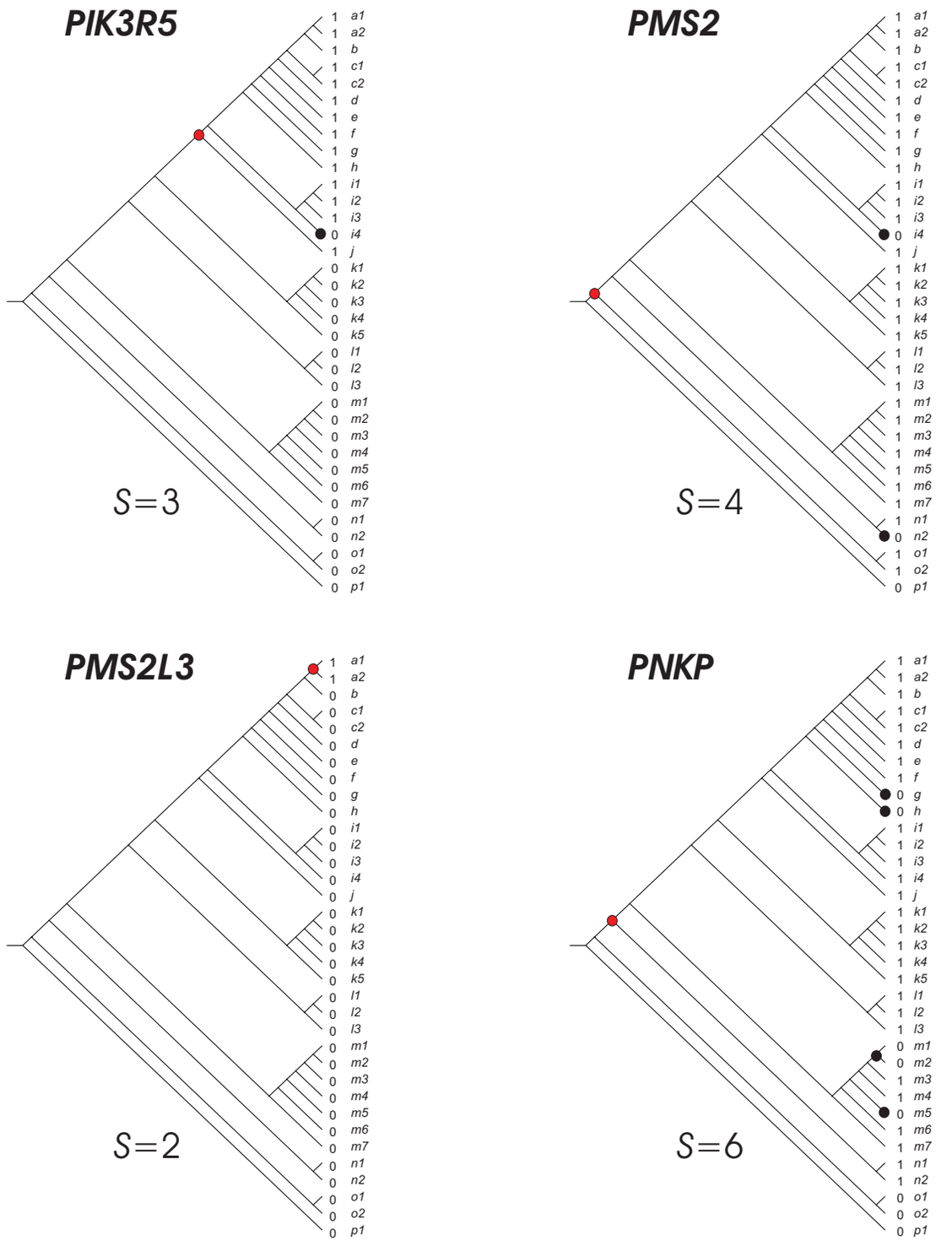
Supplementary Figure S77. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



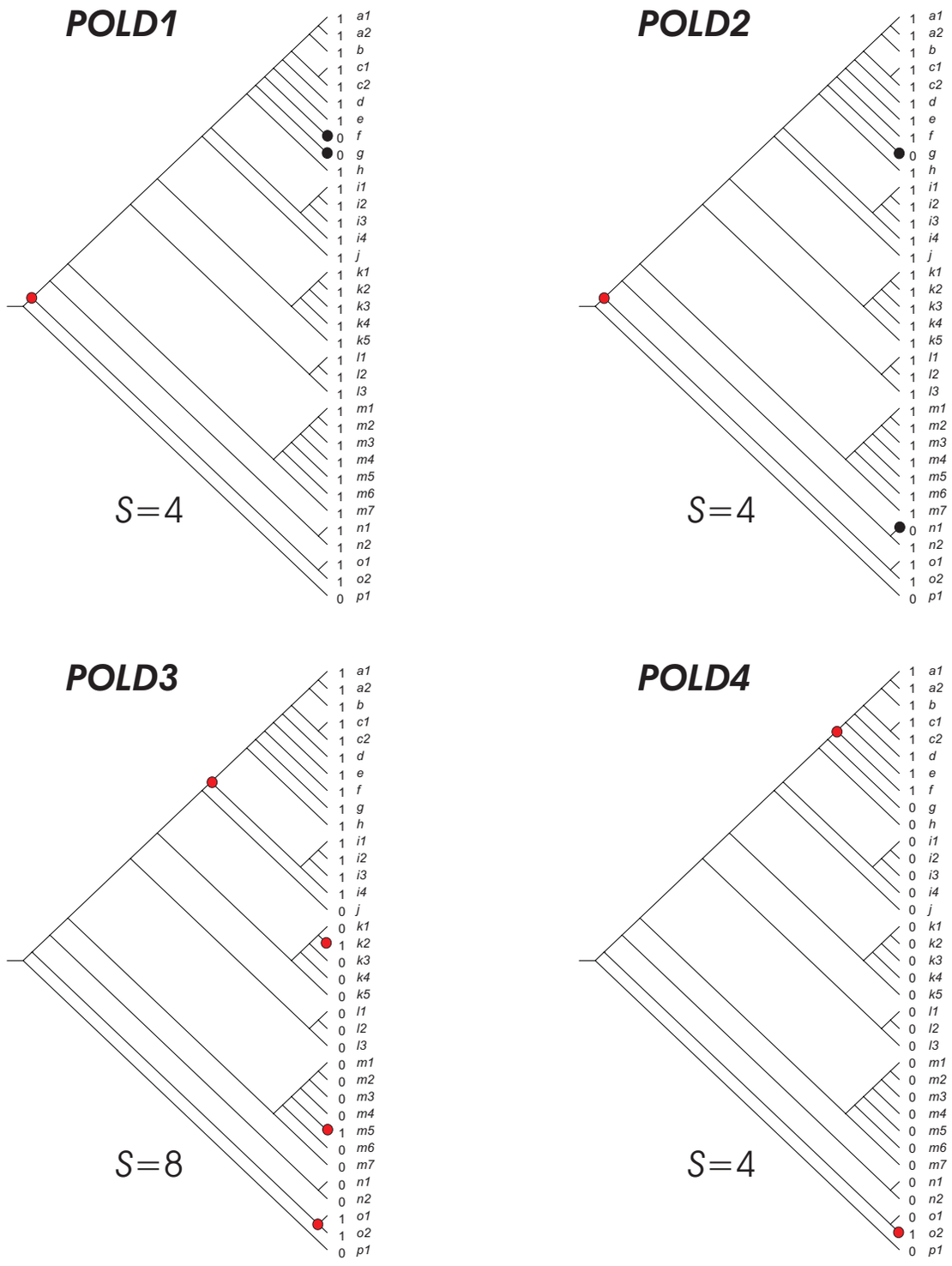
Supplementary Figure S78. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



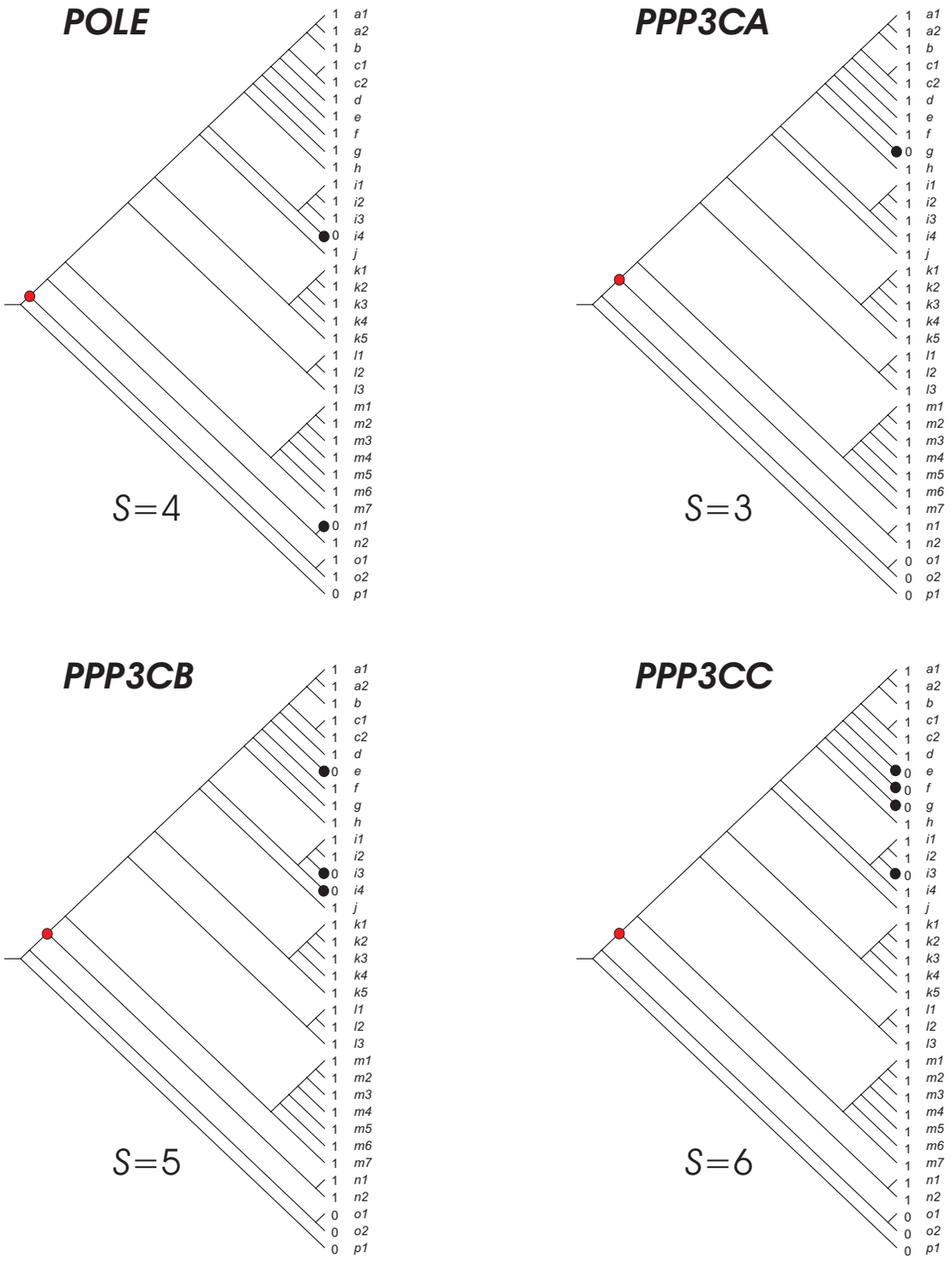
Supplementary Figure S79. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



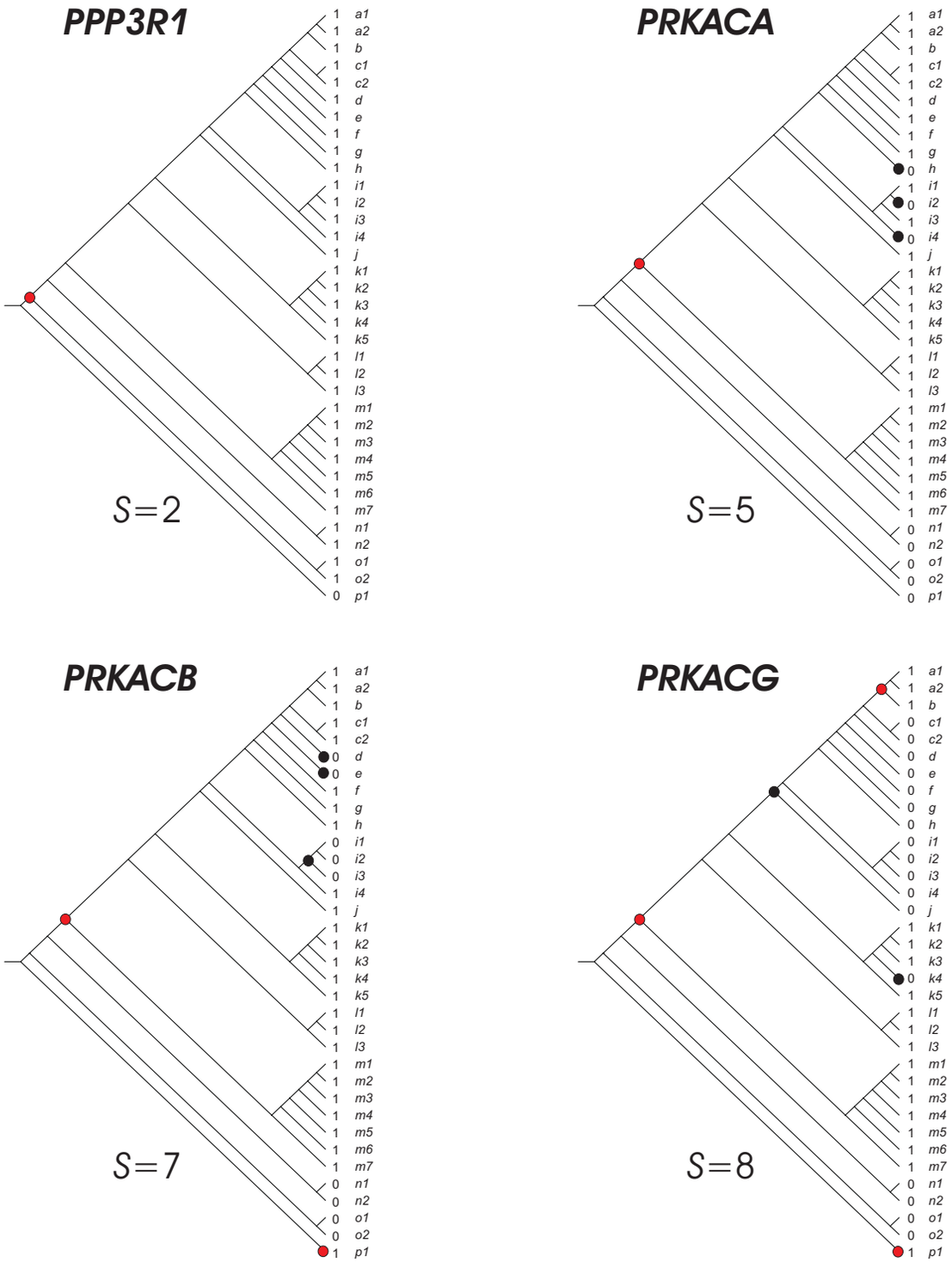
Supplementary Figure S80. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



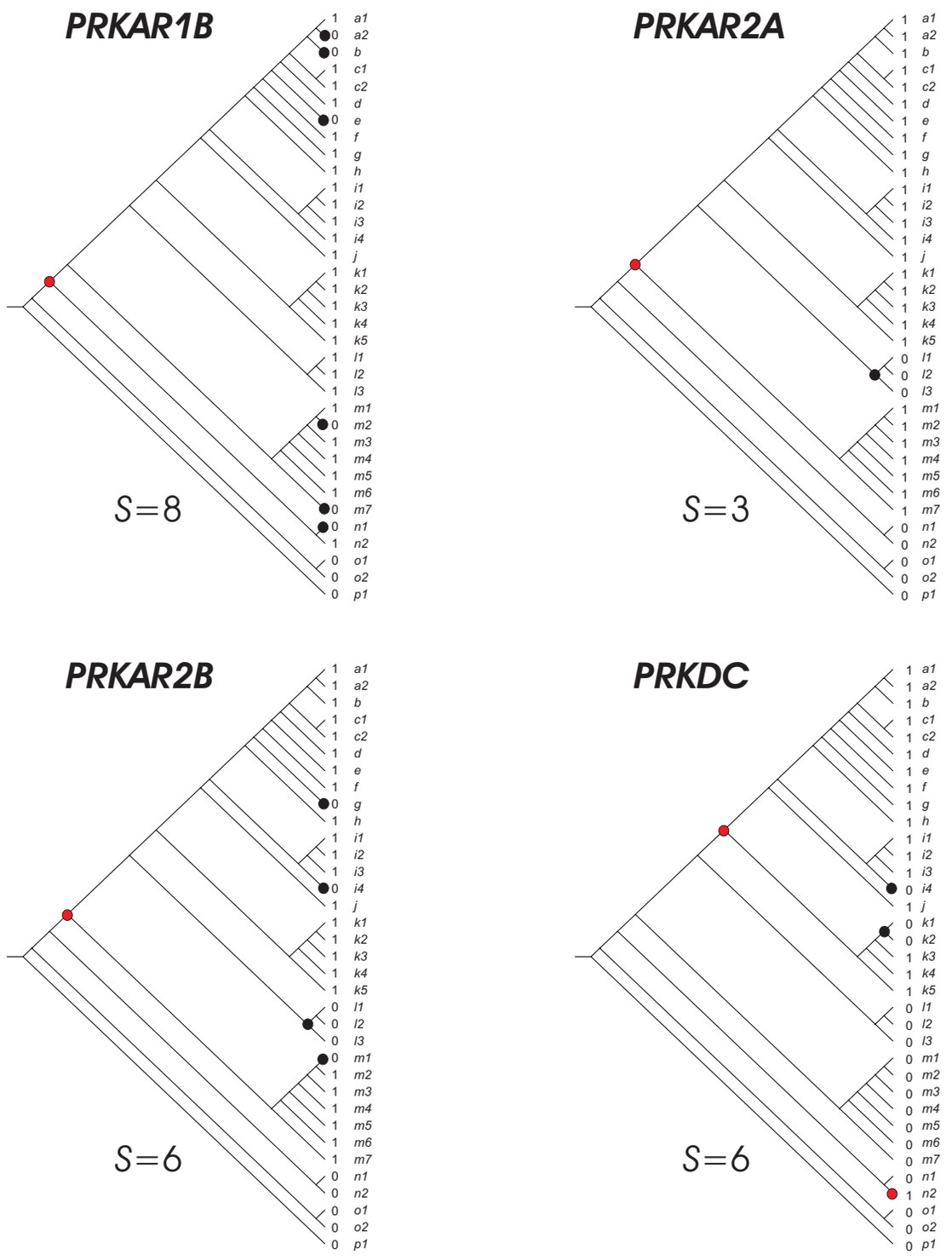
Supplementary Figure S81. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



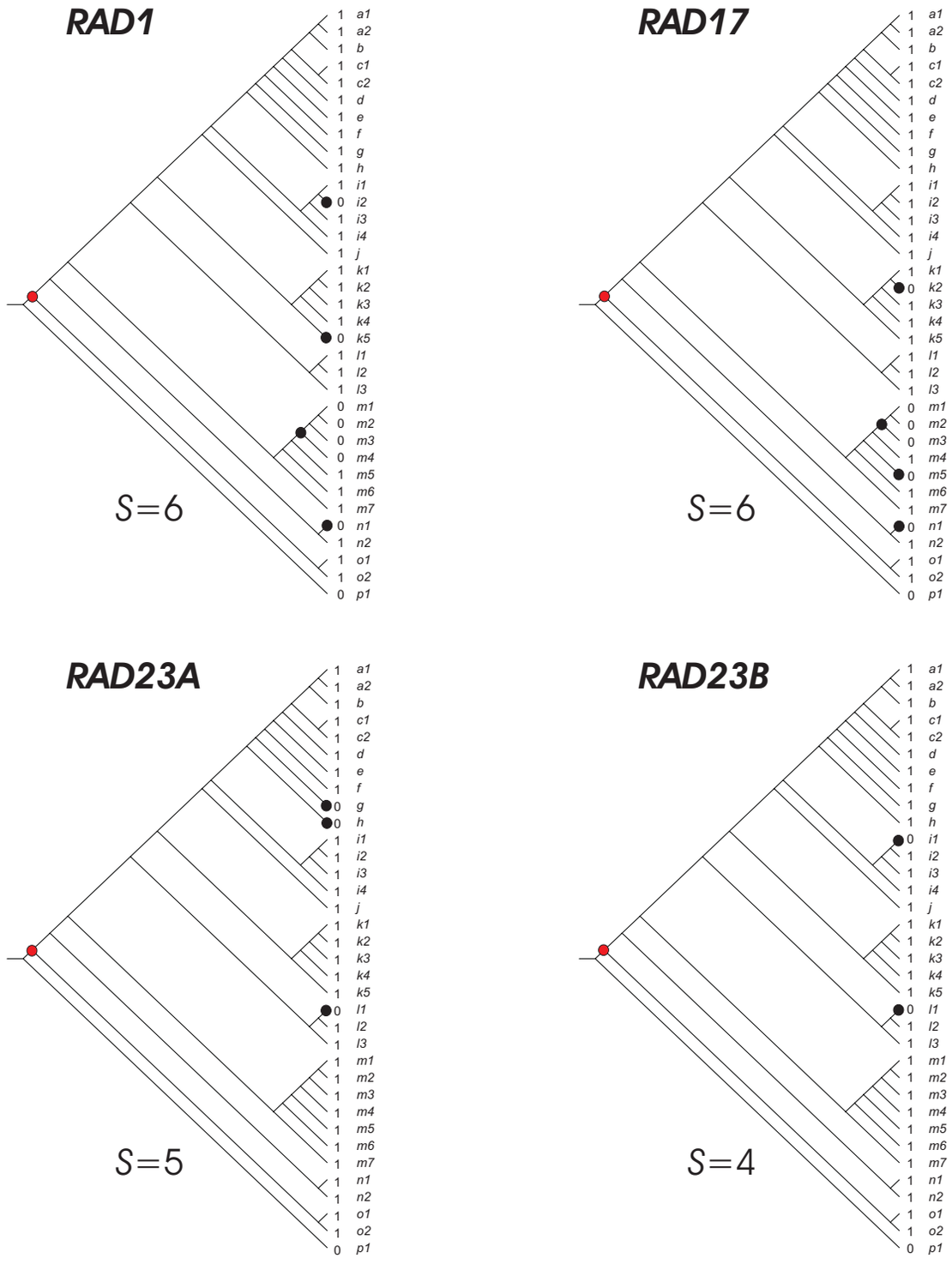
Supplementary Figure S82. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



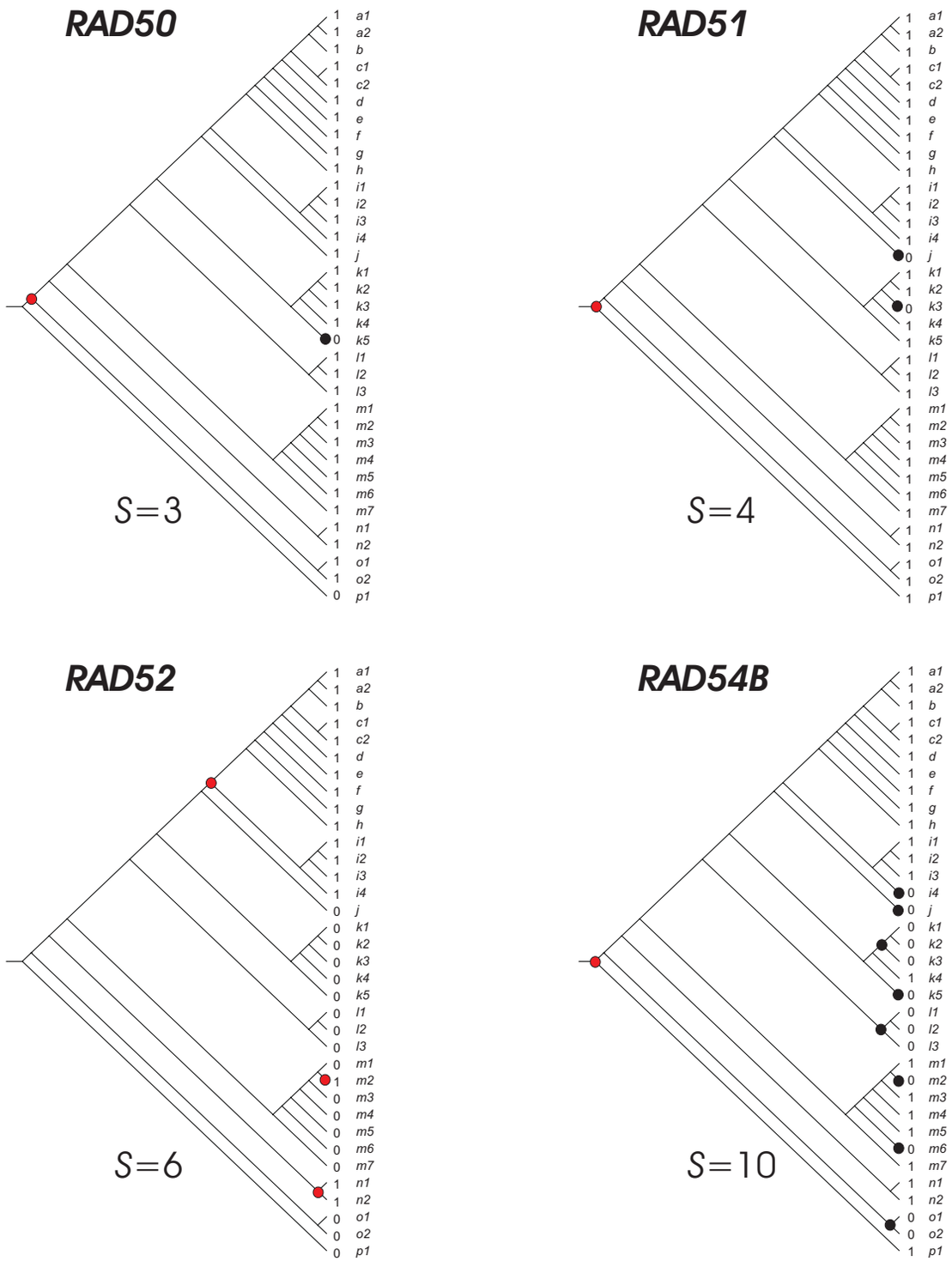
Supplementary Figure S83. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



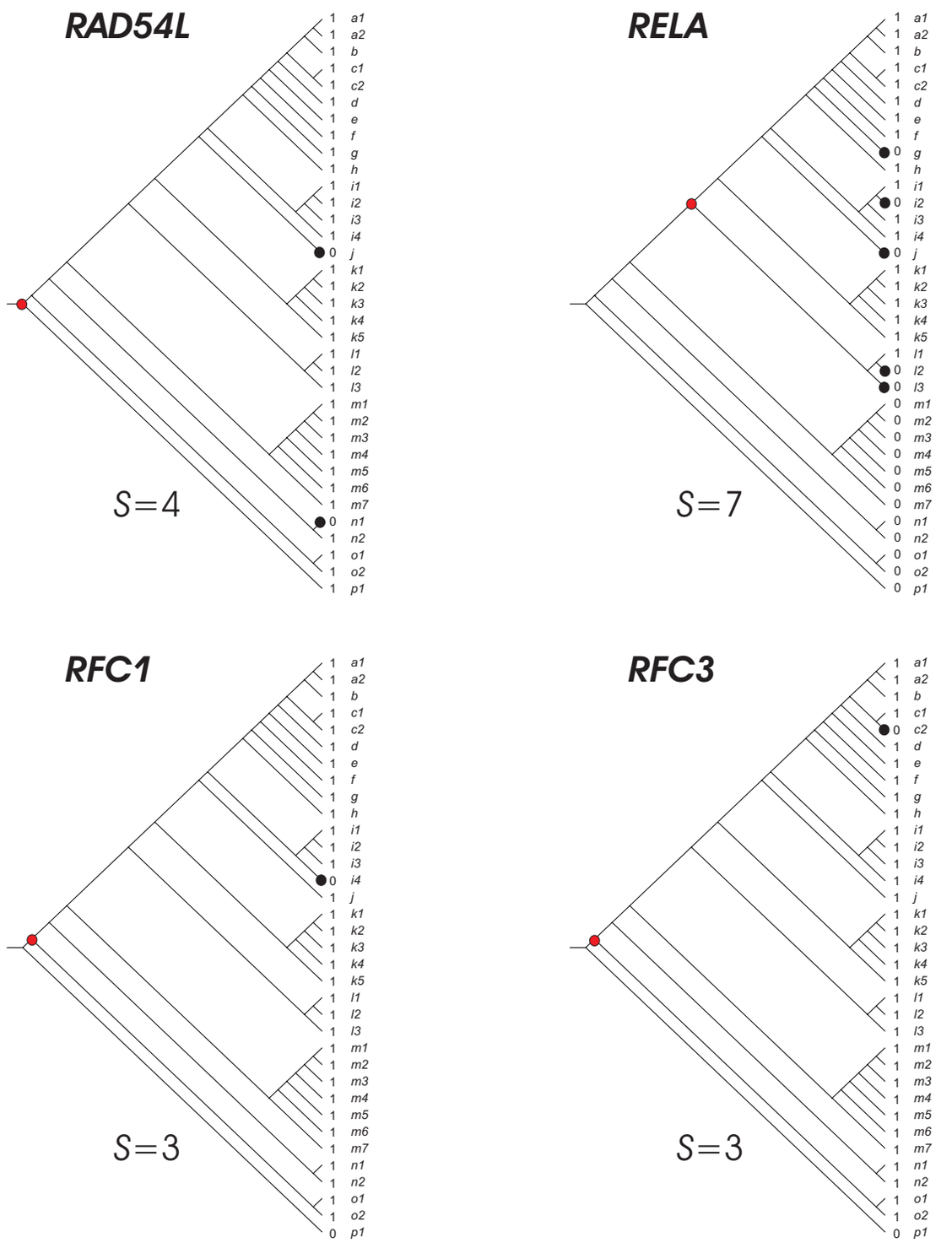
Supplementary Figure S84. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



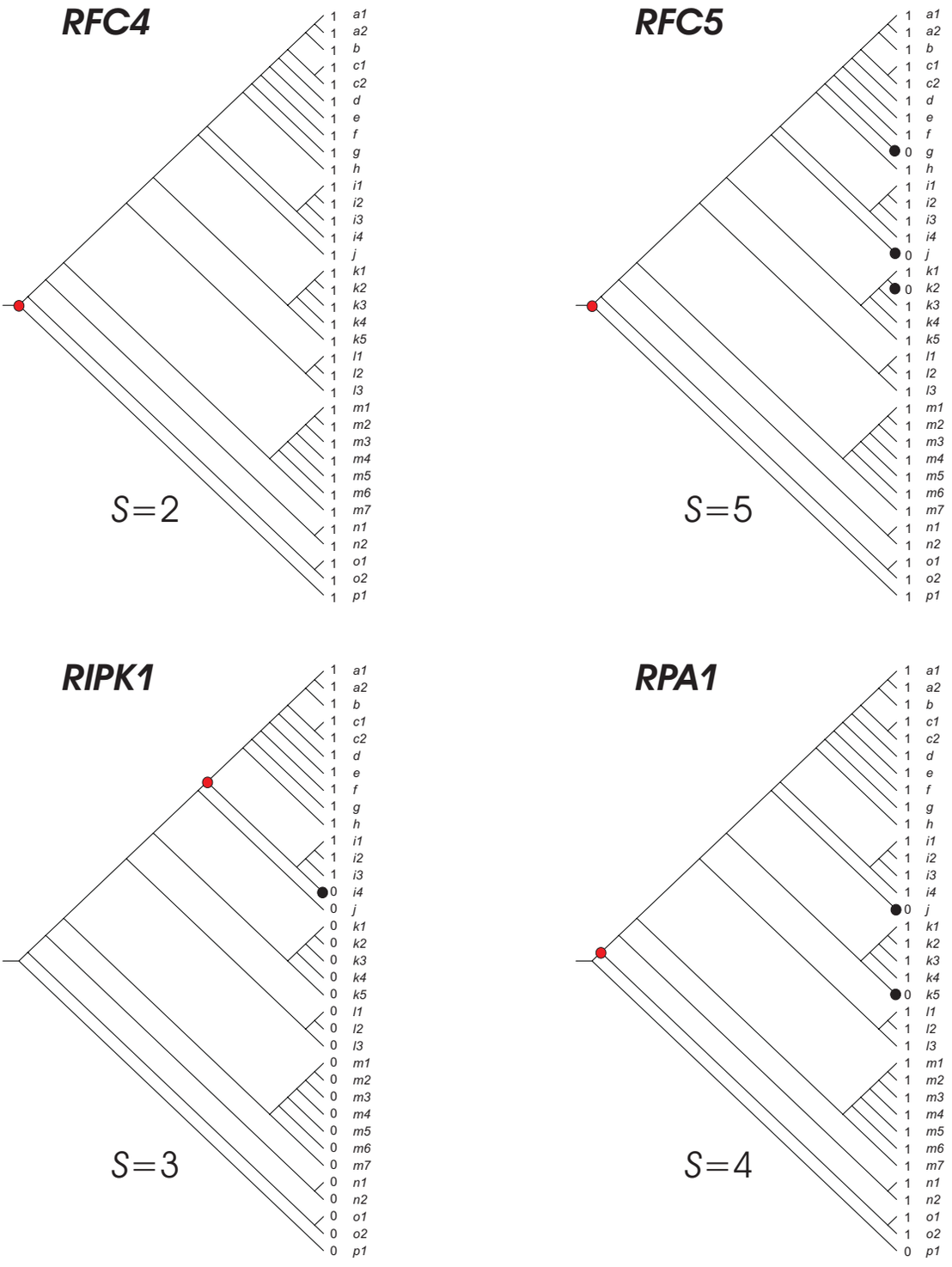
Supplementary Figure S85. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



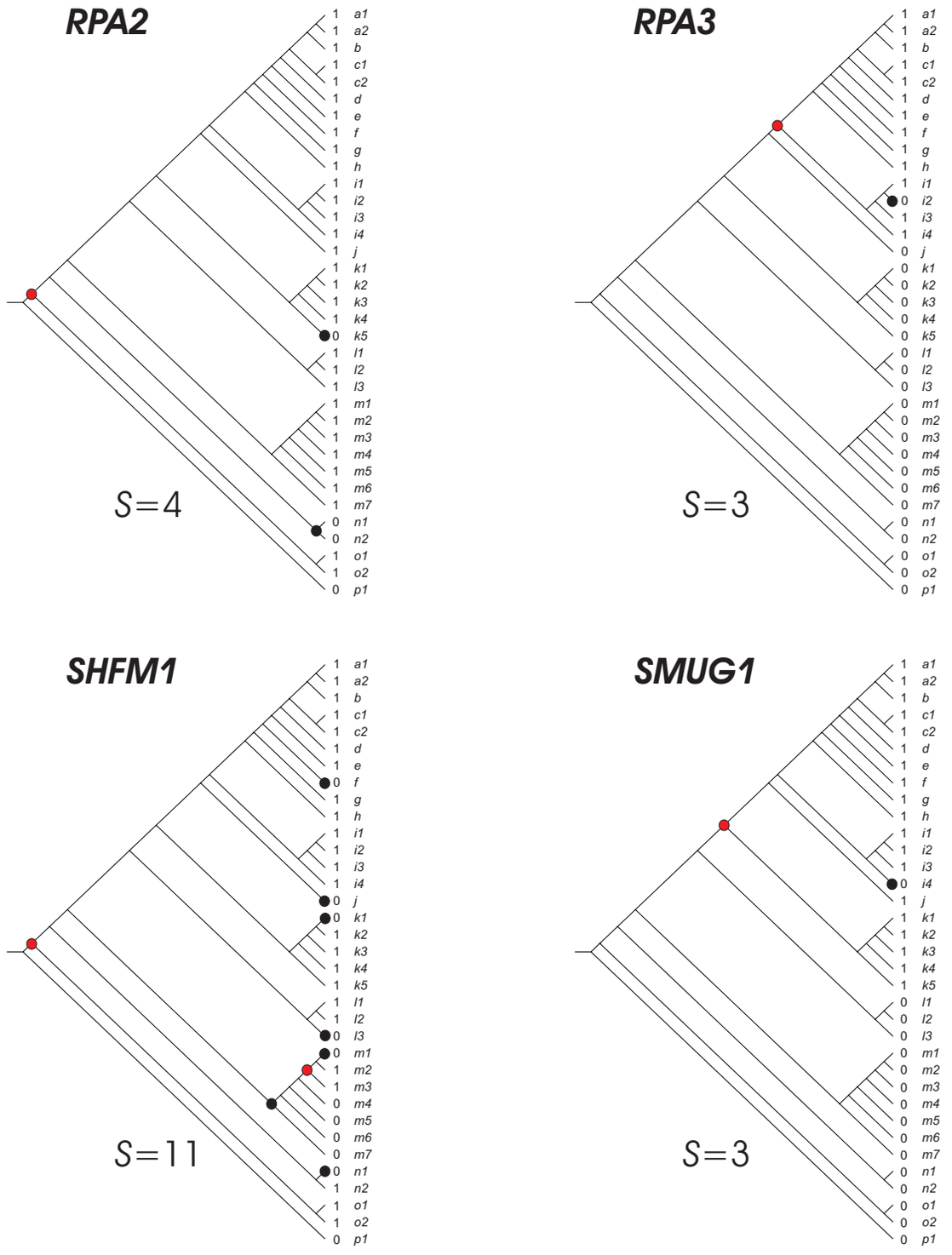
Supplementary Figure S86. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



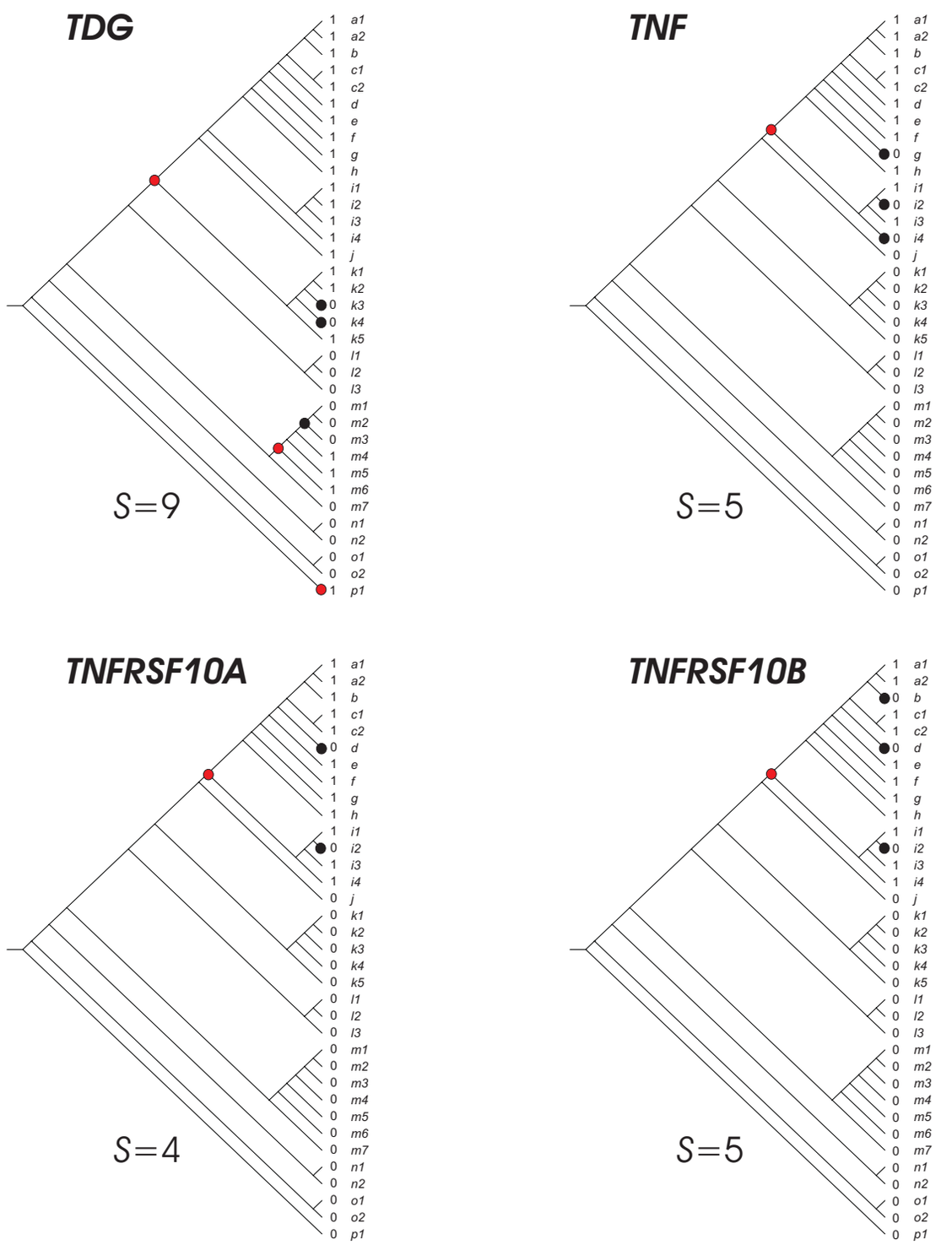
Supplementary Figure S87. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



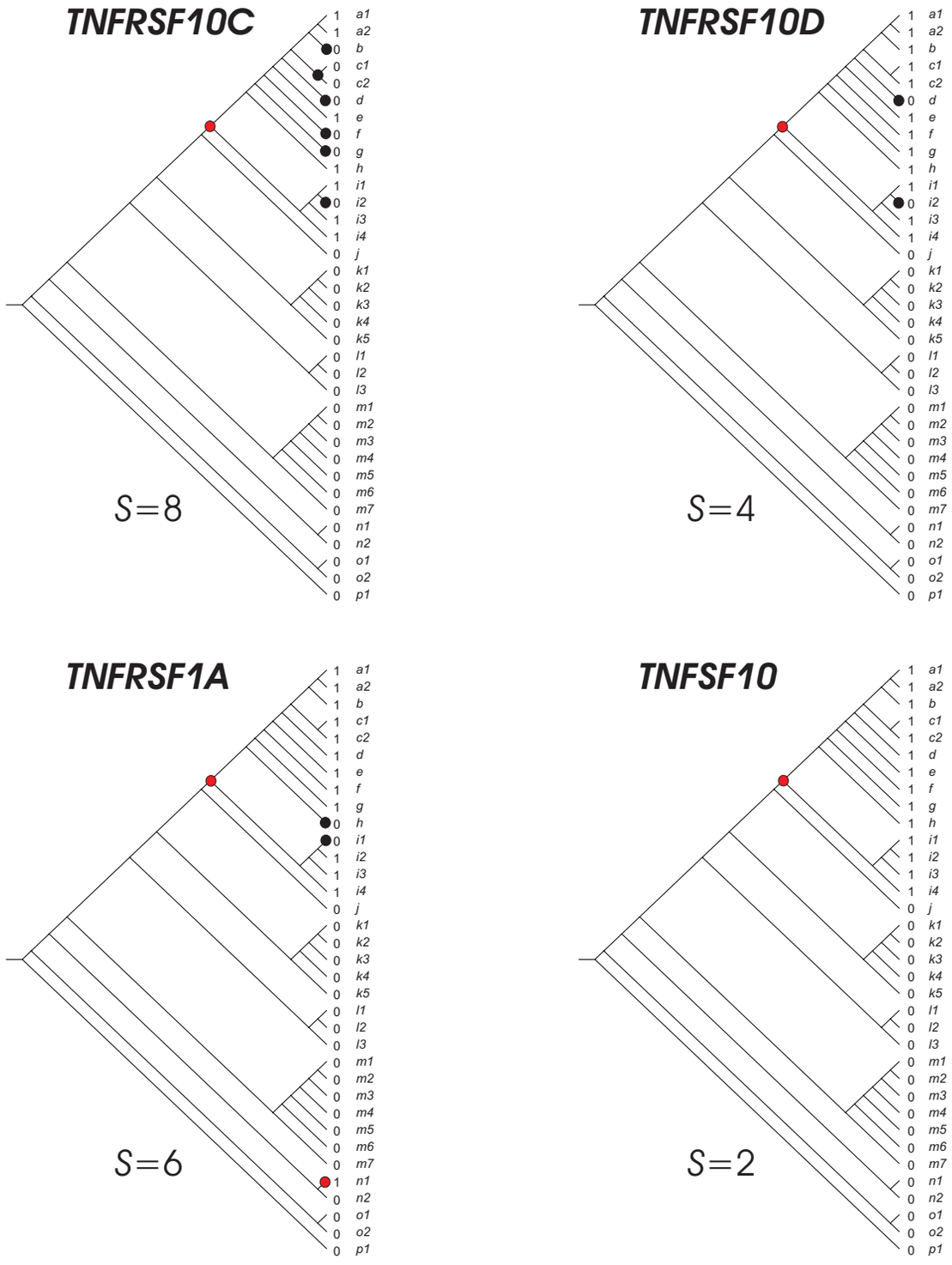
Supplementary Figure S88. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



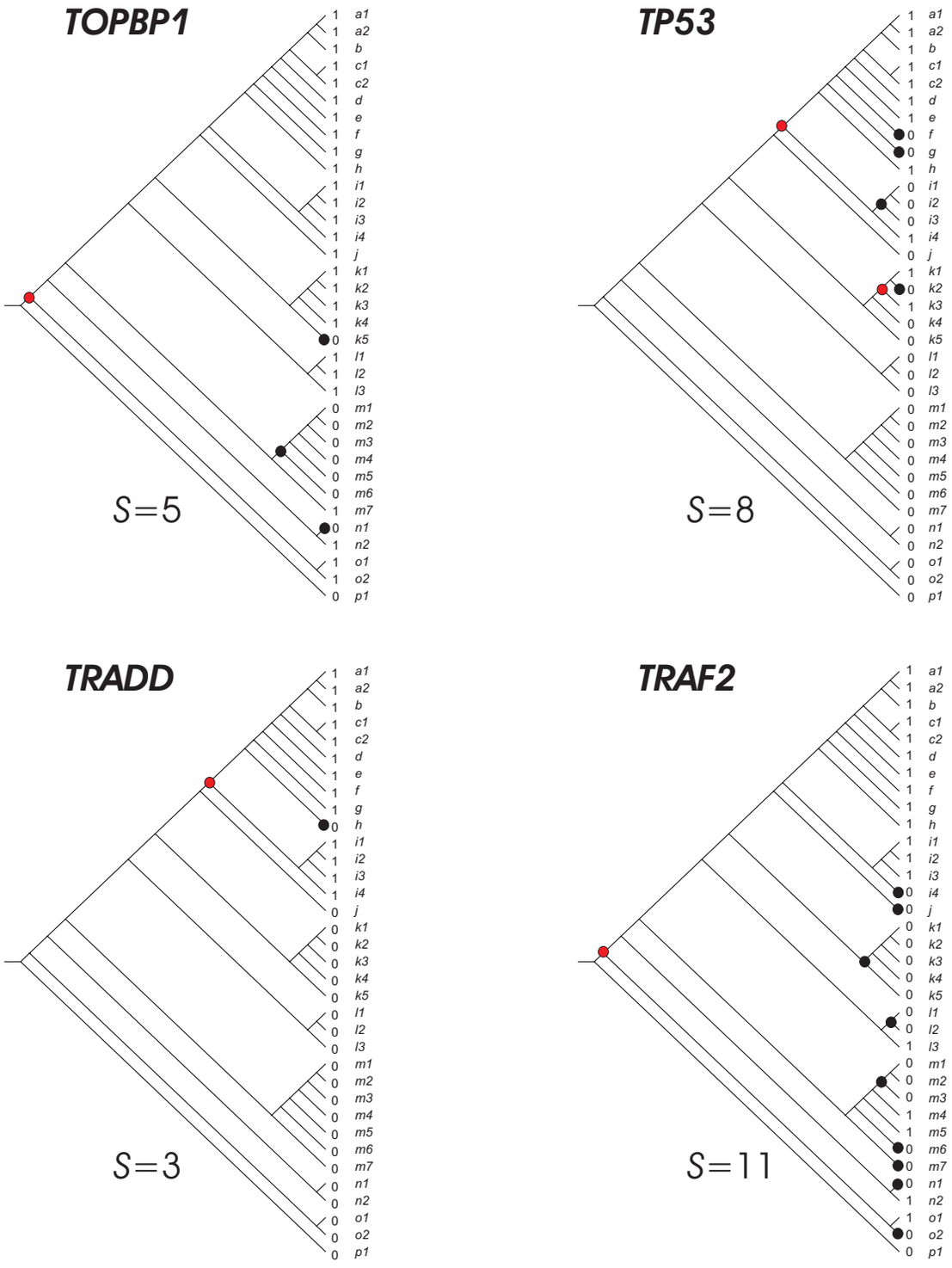
Supplementary Figure S89. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



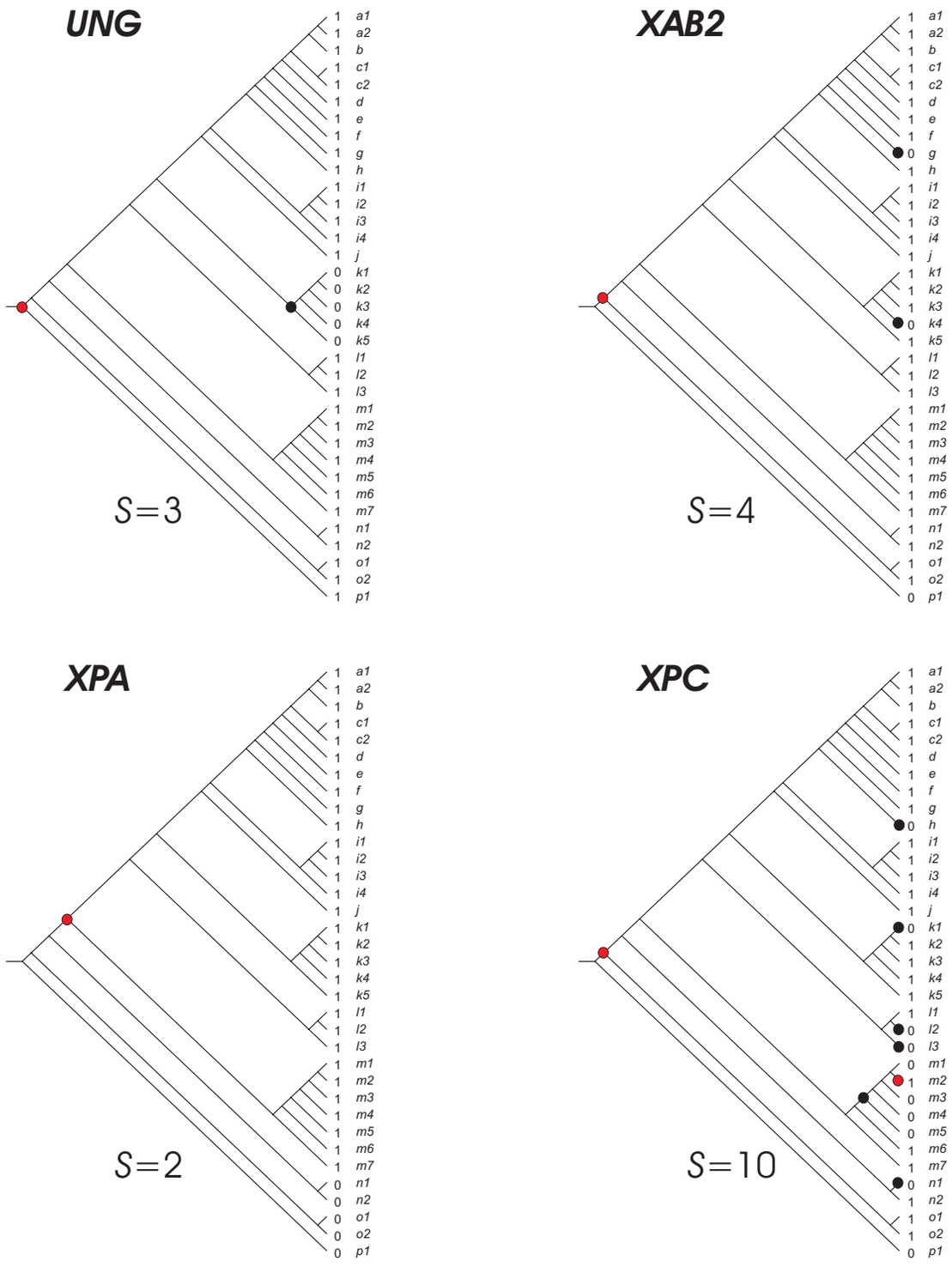
Supplementary Figure S90. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost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



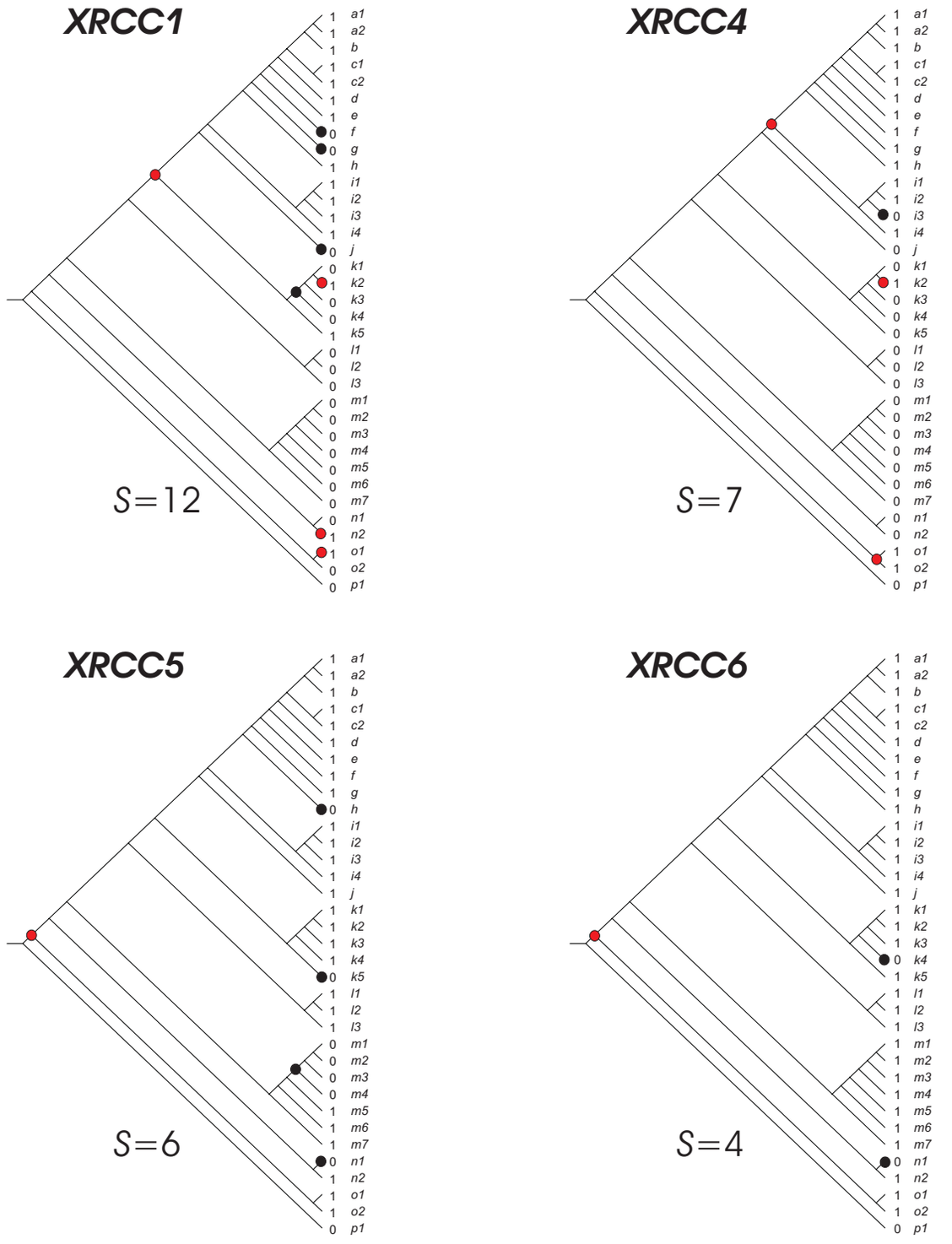
Supplementary Figure S91. Parsimony analysis of orthologous groups according to Inparanoid database. The inconsistency value *S* is indicated for each evolutionary scenario. The relative costs of the evolutionary events consider two cost units for gene birth or gene acquisition (red nodes), and one cost 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.

Evolutionary plasticity determination by orthologous groups distribution – Material suplementar.

Evolutionary plasticity determination by orthologous groups distribution

Supplementary Material

Rodrigo JS Dalmolin[1§], Mauro AA Castro [1], José L Rybarczyk Filho[2], Luis HT Souza[1]; Rita MC de Almeida[2]; José CF Moreira[1].

[1]Department of Biochemistry, Institute of Basic Health Sciences, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil.

[2]Department of Physics, Institute of Physics, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil.

Correspondent Author – Rodrigo JS Dalmolin. Departamento de Bioquímica, ICBS-UFRGS. Rua Ramiro Barcelos, 2600, anexo. Laboratório 32. Phone +55 51 33085578, fax +55 51 33085540. E-mail: rodrigo.dalmolin@ufrgs.br

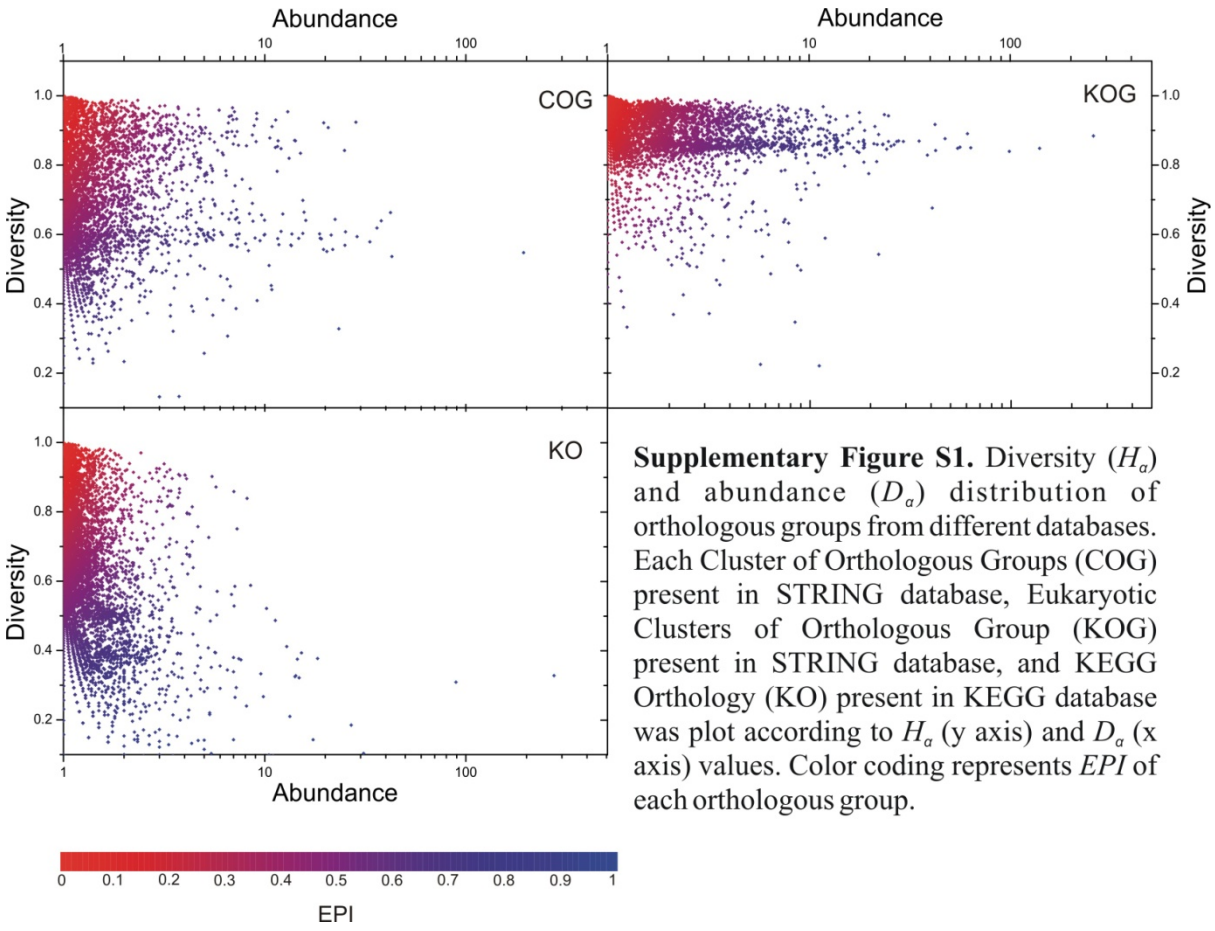
Contents

List of Supplementary Figures

1. Supplementary results and discussion

1.1. Distribution of orthologous groups from different datasets

The three most important databases concerned in organizing orthologous groups are Inparanoid Database (http://inparanoid.sbc.su.se), KO (Kegg Orthology) Database [1] (http://www.genome.jp/kegg/ko.html), and COG (Cluster of Orhologous Groups) Database (http://www.ncbi.nlm.nih.gov/COG) [2]. Inparanoid is designed to find orthologs and in-paralogs between two species and to separate in-paralogs from out-paralogs, excluding out-paralogs in orthologous groups formation [3]. Therefore, it is inappropriate to use Inparanoid orthologous groups here, since our objective is to find all proteins that have the same ancestor gene. Both COG and KO work in gathering all paralogs and orthologs that possess the same ancestral gene in the same orthologous group. We avoid working with prokaryotic genomes (*i.e.* bacteria and archaea) due to their higher proportion of horizontal gene transfer events comparing to eukaryotes [4]. Accordingly, we have used KOG dataset to perform our investigation instead of using the entire COG database (including 55 eukaryotes and 575 prokaryotes) or KO database (including 149 eukaryotes and 1164 prokaryotes). KOG extracted from STRING Database (http://string.embl.de/) represents a curated dataset to identify protein families with the same ancestral gene [5]. Supplementary Figure S1 shows abundance (D_α) and diversity (H_α) distribution of all orthologous groups from three different datasets: COG, KOG, and KO. While KOG orthologous groups present a concentrated H_α distribution (around 0.8 to 1), both COG and KO presents a wide H_α distribution, probably due to high heterogenicity of the species (*i.e.* eukarya, archaea, and bacteria) that compose each dataset.



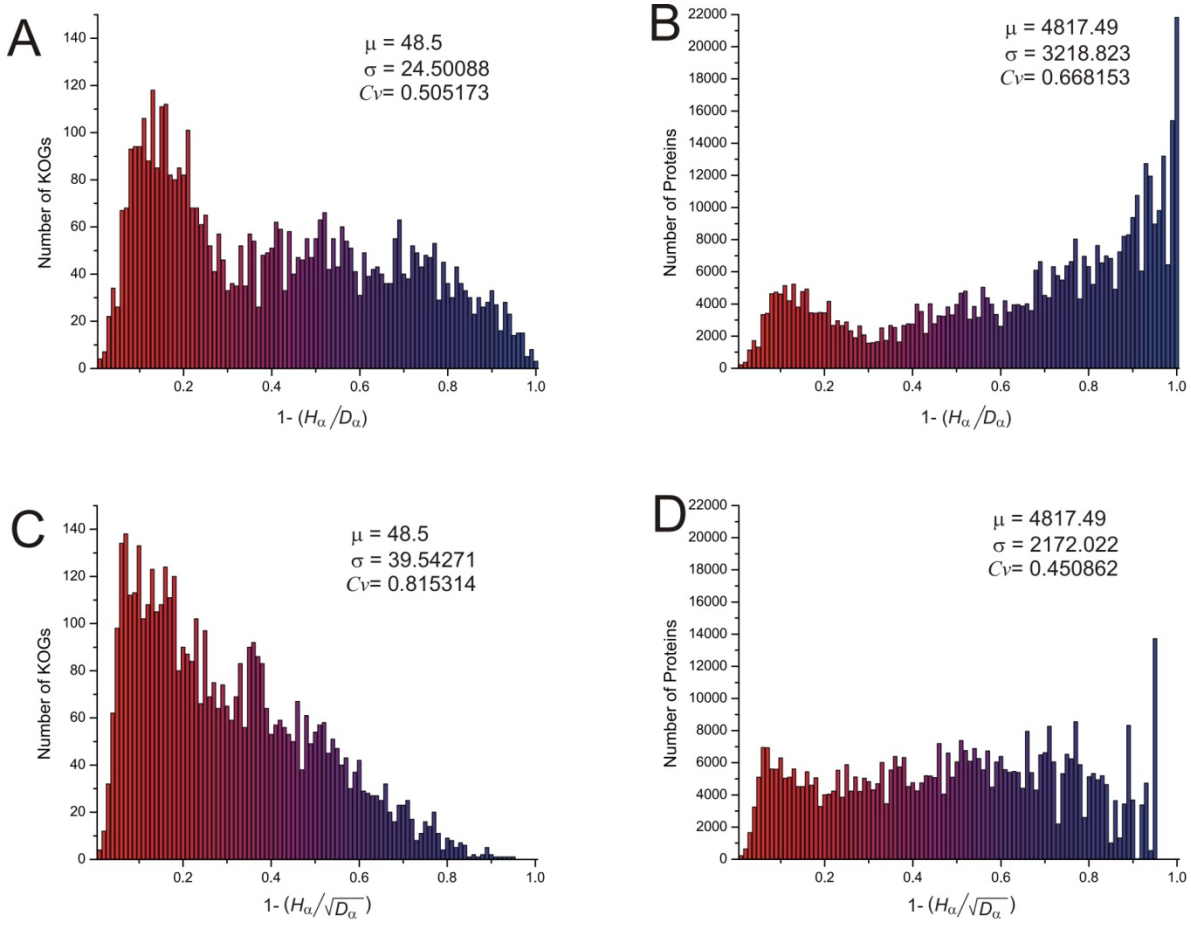
Supplementary Figure S1. Diversity (H_α) and abundance (D_α) distribution of orthologous groups from different databases. Each Cluster of Orthologous Groups (COG) present in STRING database, Eukaryotic Clusters of Orthologous Group (KOG) present in STRING database, and KEGG Orthology (KO) present in KEGG database was plot according to H_α (y axis) and D_α (x axis) values. Color coding represents *EPI* of each orthologous group.

1.2. *EPI* equation determination

According to what was discussed in the main text, evolutionary plasticity of an orthologous group is positively correlated to their D_α and negatively correlated to their H_α. The most intuitive way to produce an index comprising diversity and abundance together is the ratio between both. However, not necessarily the ratio between H_α and D_α will best represent the evolutionary plasticity of an orthologous group. Supplementary Figure S2 shows all KOGs present in STRING (Supplementary Figure S2.A and S2.C) and all proteins that compose those KOGs (Supplementary Figure S2.B and S2.D). KOGs and proteins were organized according to evolutionary plasticity as follows: equation (1) (Supplementary Figure S2.A and S2.B) and equation (2) (Supplementary Figure S2.C and S2.D).

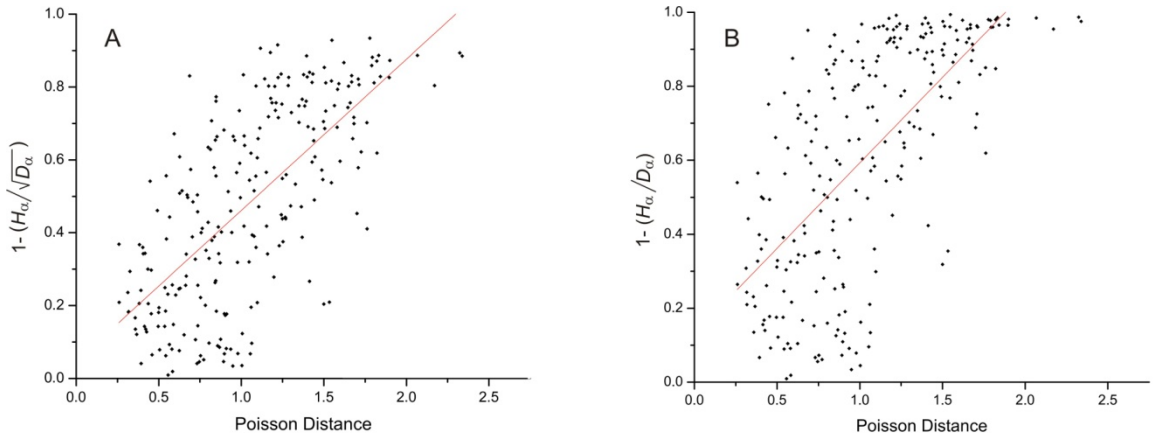
$$Evolutionary\ Plasticity = 1 - \frac{H_\alpha}{D_\alpha} \quad (1)$$

$$Evolutionary\ Plasticity = 1 - \frac{H_\alpha}{\sqrt{D_\alpha}} \quad (2)$$



Supplementary Figure S2. Evolutionary Plasticity distribution according to different equations. All KOGs and proteins present in STRING database were grouped in 100 categories according to evolutionary plasticity calculated according to equation (1) (A and B, respectively) and equation (2) (C and D, respectively). Color coding is proportional to evolutionary plasticity. μ = mean, σ = standard deviation, and *Cv* = coefficient of variation.

The distribution of all proteins present in KOG dataset was dislocated to high plasticity when equation (1) was used to determine evolutionary plasticity (Supplemental Figure S2.B). Conversely, we observed an equalized protein distribution according to evolutionary plasticity when using equation (2) (Supplemental Figure S2.D). Thus, a protein randomly chosen among all species present in KOG dataset has similar probability to show low, median, or high evolutionary plasticity using equation (2) to determine evolutionary plasticity. Following those criteria, equation (2) has been elected to describe the Evolutionary Plasticity Index (*EPI*). To certify the competence of equation (2) comparing to equation (1), we repeated evolutionary distance analysis and functional plasticity analysis using both equations to determine evolutionary plasticity. Both analyses have been performed with the same methodology described in *Material and Methods* section (*Molecular Evolutionary Analysis* and *Fitness Evaluation*), except when changing equation (2) by equation (1) in *EPI* determination. Supplementary Figure S3 shows two correlation graphics between evolutionary distances among all proteins present in a same KOG and evolutionary plasticity of that KOG calculated by equation (2) (Supplementary Figure S3.A) and by equation (1) (Supplementary Figure S3.B).



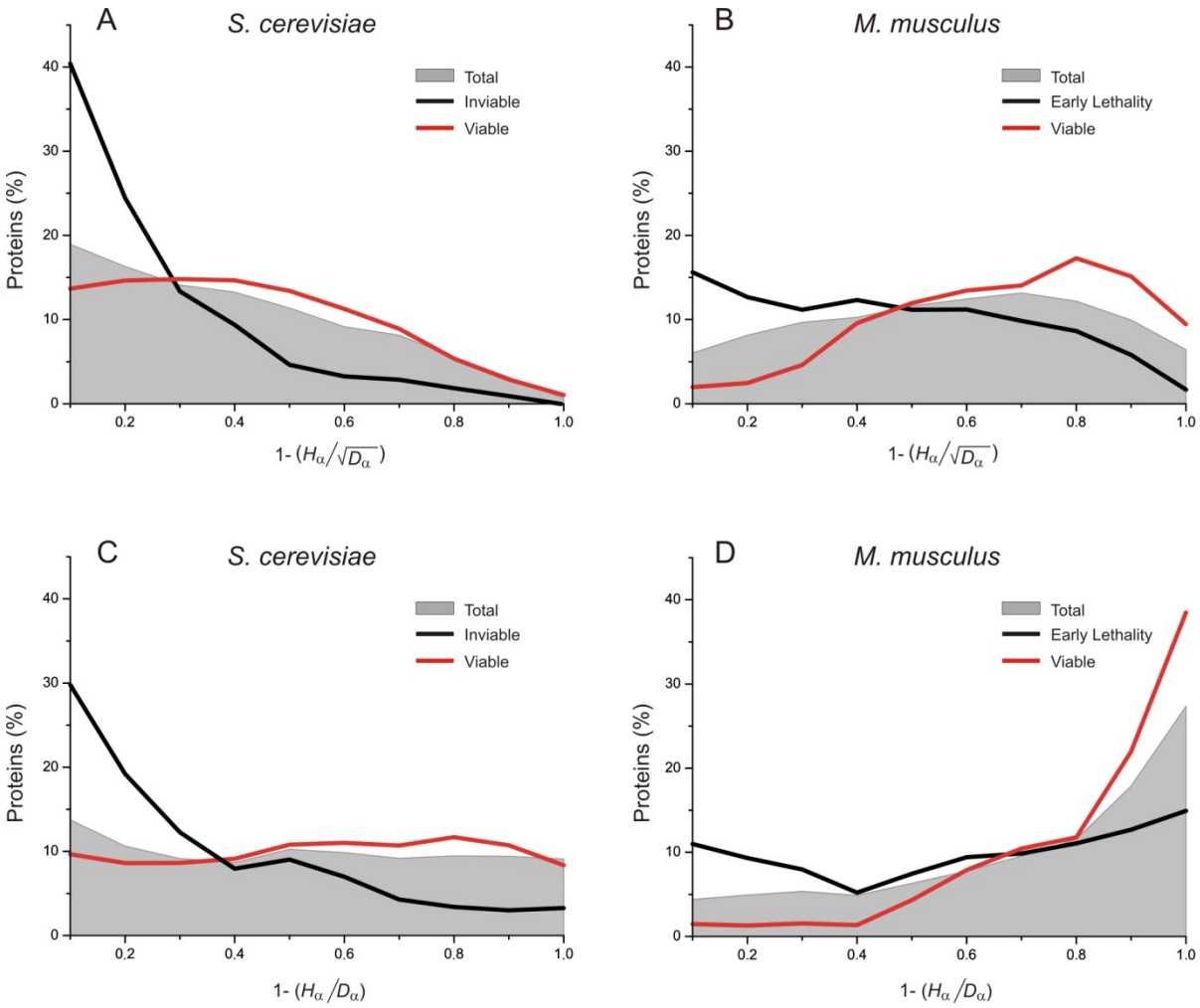
Supplementary Figure S3. Evolutionary Distance Average versus Evolutionary Plasticity according to different equations. 5% of the KOGs present in STRING database were sorted. The evolutionary distance among all proteins of each KOG evaluated was calculated and the evolutionary distance average (Poisson Distance) was obtained. Poisson Distance was plotted against evolutionary plasticity calculated according to equation (2) (A) and according to equation (1) B. Red lines indicates the linear regression fitting curve.

Supplementary Table S1 shows Pearson Correlation, as well as fitting curve properties, of both graphics. Despite both graphics have shown a correlation between evolutionary distance and evolutionary plasticity, the graphic generated using equation (2) (*i.e.* graphic A) presented higher Pearson Correlation. In addition, all linear regression fitting curve properties, such as intercept, slope, and residual sum squares, was more adequate in graphic A comparing to graphic B (Supplementary Table S1), reinforcing equation (2) utilization.

Supplementary Table S1. Linear regression fitting curve properties

	Graphic A		Graphic B	
Pearson Correlation	0.68621		0.66268	
Adj. R-Square	0.46869		0.43682	
Residual Sum of Squares	9.17866		12.93475	
	Value	Standard Error	Value	Standard Error
Intercept	0.046	0.03213	0.13076	0.03815
Slope	0.41511	0.02834	0.46222	0.03365

Supplementary Figure S4 shows the distribution of proteins from *S. cerevisiae* (Supplementary Figure S4.A and S4.C) and *M. musculus* (Supplementary Figure S4.B and S4.D) according to *EPI* calculated by equation (2) (Supplementary Figure S4.A and S4.B) and equation (1) (Supplementary Figure S4.C and S4.D). As shown in Supplementary Figure S4, the same phenomenon observed when using equation (2) to calculate evolutionary plasticity (as presented in *Results*, section *Functional Plasticity Analysis*) can be observed when using equation (1). In fact, an improvement in the differences of the means can be observed using equation (1) when compared to equation (2) (table S2 and table S3). However, except by differences in Z-value of *Saccharomyces cerevisiae* viable group, the differences are not outstanding.



Supplementary Figure S4. Distribution of target genes with different fitness impact according to Evolutionary Plasticity calculated using different equations. The percentage of *S. cerevisiae* and *M. musculus* genes presenting different evolutionary plasticity values calculated using equation (2) (A and B) and equation (1) (C and D). The grey landscape represents the *EPI* distribution of all genes from each species. Black lines represent the *EPI* distribution of *S. cerevisiae* genes associated with inviable phenotype when knocked-out (A and C) and *M. musculus* target genes associated with early lethality (B and D). Red lines represent the *EPI* distribution of target genes associated with viable phenotypes (A, B, C, and D).

Supplementary Table S2. Descriptive statistics of *S. cerevisiae* genes.

	Number of Proteins	Mean	Standard Deviation	Standard Error	|Z-value|
Equation (1)					
Total	3998	0.47477327	0.292309	-	-
Inviable	891	0.29451211	0.2503	0.00979272	18.40767312
Viable	2792	0.51078711	0.2793	0.00553203	6.510059513
Equation (2)					
Total	3998	0.3430	0.2334	-	-
Inviable	891	0.2010	0.1832	0.0078	18.1569
Viable	2792	0.3690	0.2242	0.0044	5.8988

Supplementary Table S3. Descriptive statistics of *M. musculus* genes.

	Number of Proteins	Mean	Standard Deviation	Standard Error	|Z-value|
Equation (1)					
Total	14919	0.6726	0.2821	-	-
Early lethality	368	0.5474	0.3076	0.0147	8.5165
Viable	244	0.7858	0.2164	0.0181	6.2674
Equation (2)					
Total	14919	0.5199	0.2565	-	-
Early lethality	368	0.4061	0.2561	0.0134	8.5169
Viable	244	0.6226	0.2211	0.0164	6.2501

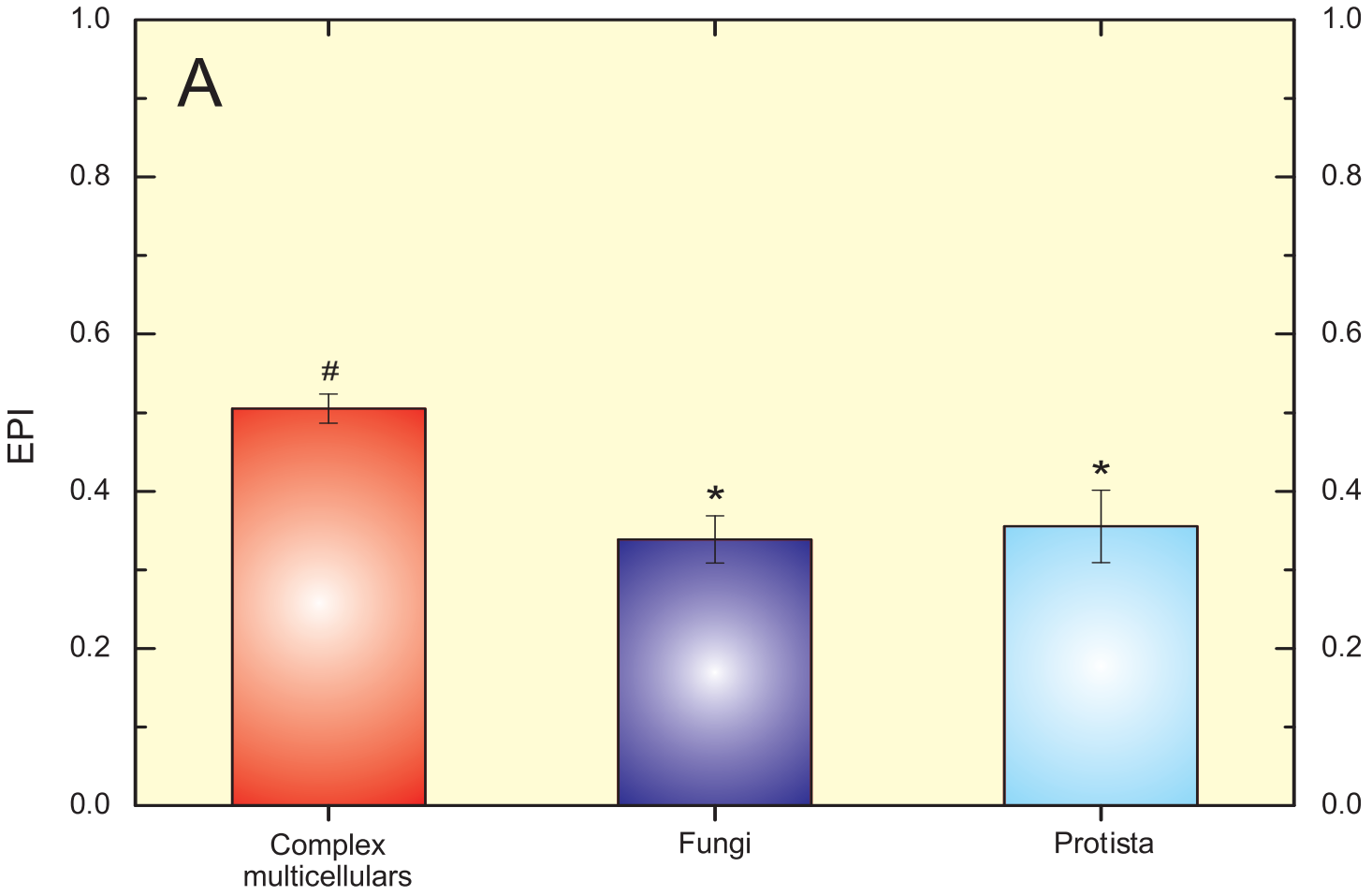
1.3. Lethality information

Several studies concerning essentiality and evolutionary parameters (*e.g.* duplicability and evolutionary rate) have been performed, hardly ever present apparent conflicting results [6-9]. However, care must be taking to compare unicellular and multicellular organisms according to lethality. Commonly, a mammalian gene is considered lethal when its deletion leads to organism death in any phase of development, in the first moments after birth, or even by causing infertility. Following those criteria, one will find essential genes which have arrived in different moments of evolution, turning difficult to trace a relationship among any evolutionary parameter and lethality in mammals. Increase in complexity is a hallmark of life [10] and the impairment of any organizational level can be lethal to complex organisms. However, impairment in biological systems which have arrived early in evolution (*i.e.* before multicellularity) might lead to early developmental lethality. In other words, a system such as DNA repair is important to unicellular organisms and a disruption in its homeostasis can be lethal to the cell, leading to early lethality. Systems involved in maintaining tissue homeostasis (*e.g.* apoptosis) may lead to lethality, however in earlier development stages compared to DNA repair, since a single cell con survive without apoptosis [11]. In a recent paper, Chen and colleagues have investigated the effect of young genes deletion in *D.*

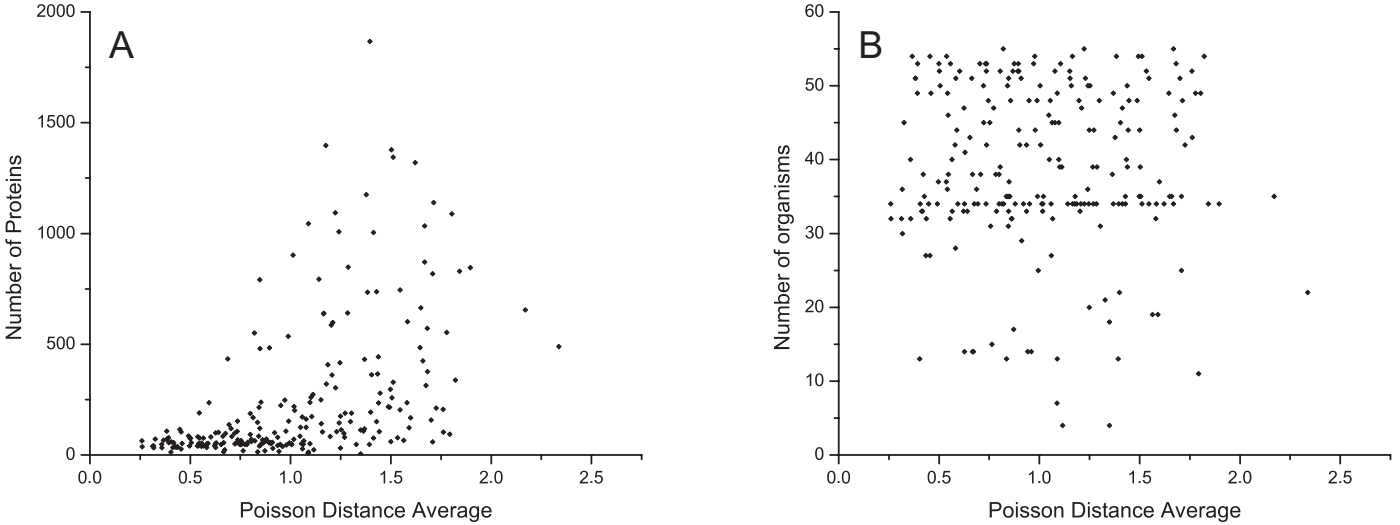
melanogaster. They observed lethality associated with young genes mainly in middle or late stages of development. Indeed, no one deleted young gene lead to lethal phenotype at first or second larval instar [12]. Those results agree with the idea that young genes have less probability to lead to early lethality.

1.4. *EPI* distribution in different species

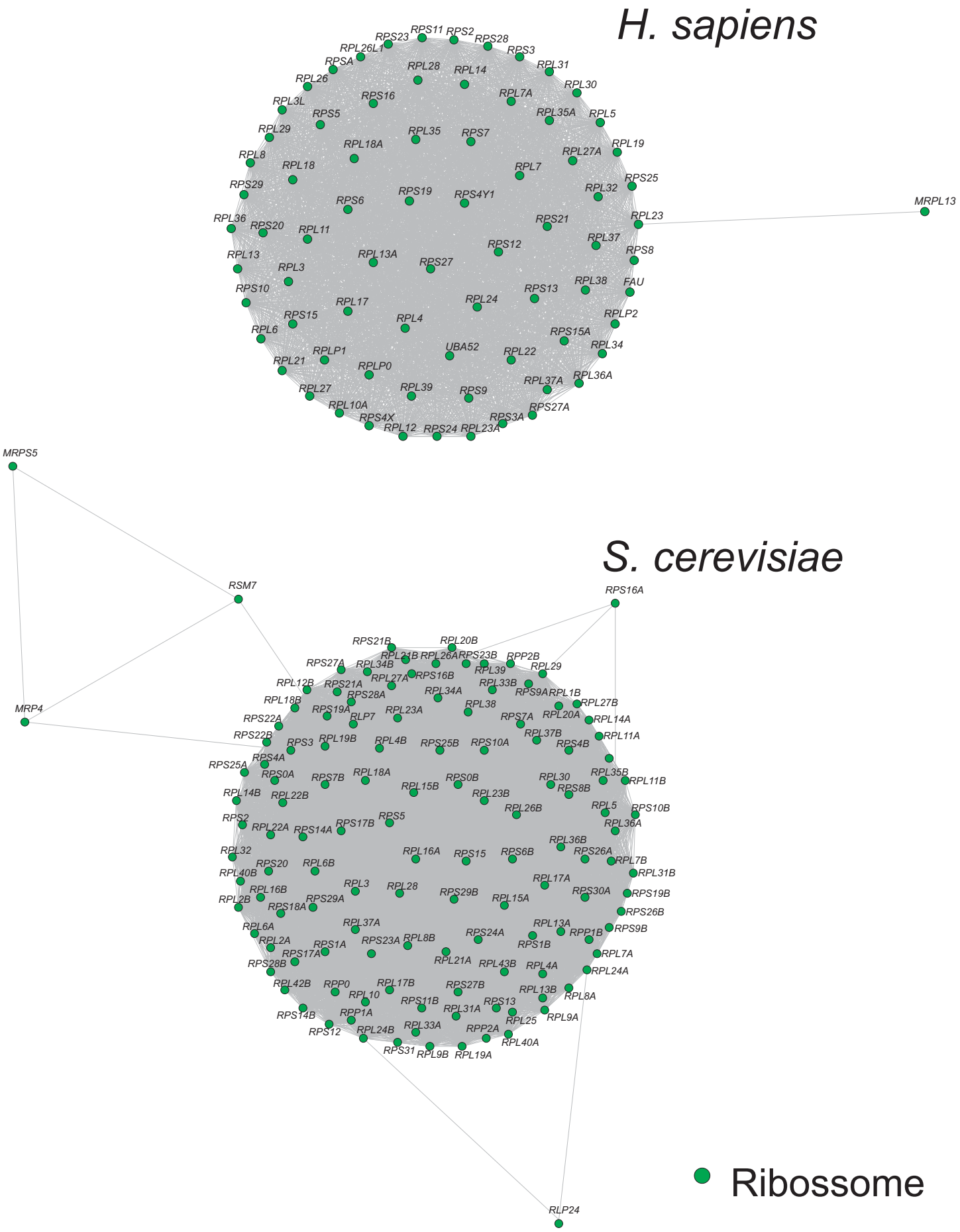
According to what was discussed in main text, simple organisms have a great proportion of low *EPI* genes comparing to complex organisms. Supplementary Figure S5 shows *EPI* distribution of all proteins present in KOG dataset from different taxonomic groups (complex multicellulars, fungi, and protista). There is a similar distribution between complex multicellular organisms, as well as between fungi and protista (Figure S5 and Figures S10 to S23). While the average *EPI* of proteins from complex multicellular organisms is around 0.5, average *EPI* of proteins from simple organisms is around 0.35. Supplementary Figure S5 considers average *EPI* from each species (*i.e.* average *EPI* from all proteins from a given species) from complex multicellular organisms (*i.e.* metazoa and plantae merged), fungi, and protista. While *EPI* from fungi and protista species do not significantly differed among each other, *EPI* from the species of both groups are significantly lower than *EPI* from complex multicellular species. Multicellular organisms possess genes that have appeared in different moments of evolution. For example, a great number of genes responsible to cellular homeostasis arrived before multicellularity advent (*e.g.* DNA repair genes [13]), whereas a great proportion of the genes involved in cell-cell communication arrived during multicellular evolution (*e.g.* TNF family [14]). However, unicellular organisms might possess higher proportion of ancient conserved genes when compared to mammals or other multicellular organisms.



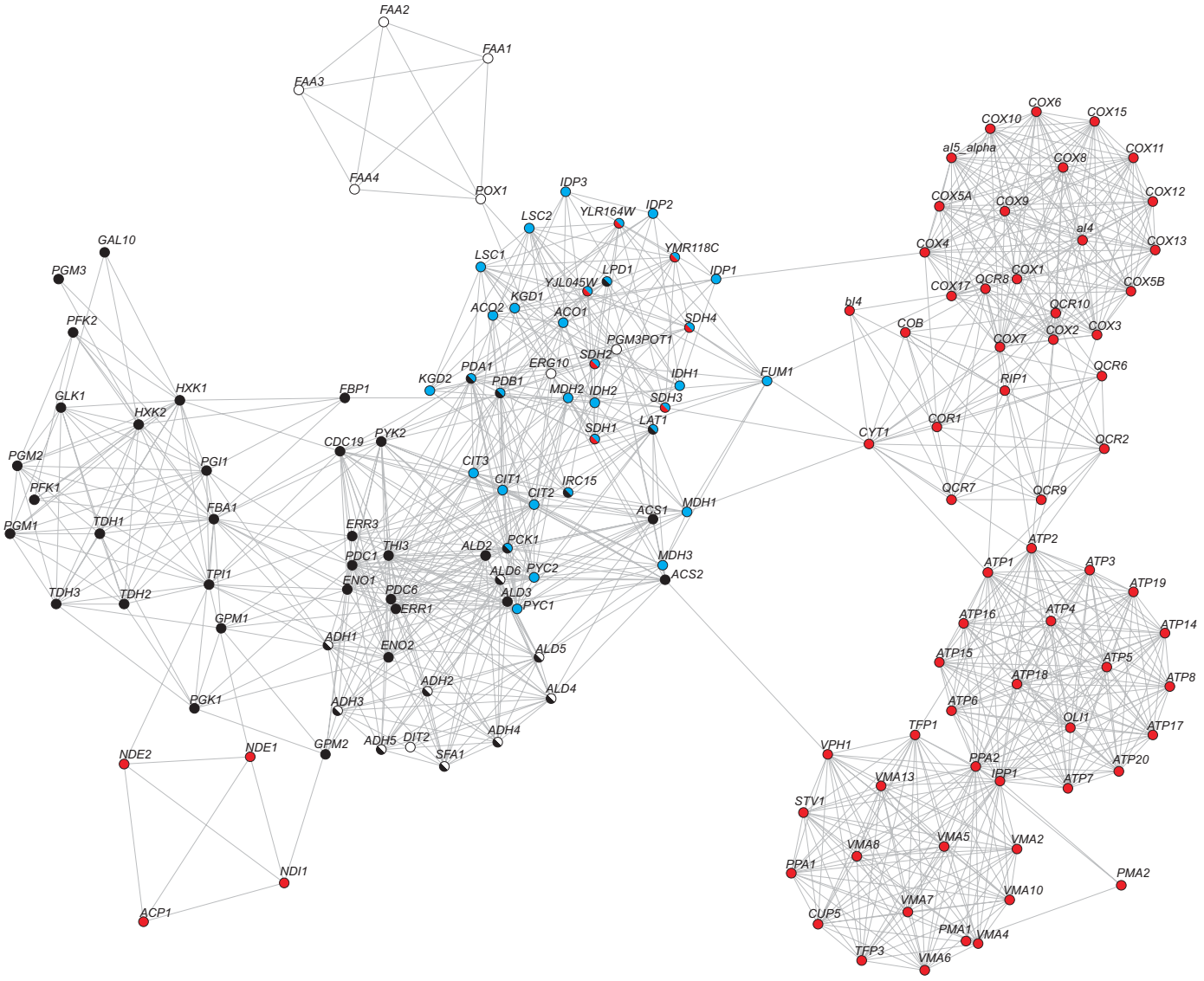
Supplementary Figure S5. *EPI* distribution of different groups according to complexity. Proteins of different groups were considered together to identify *EPI* distribution of the respective group. Complex multicellular organisms: *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Otolemur garnettii*, *Tupaia belangeri*, *Mus musculus*, *Rattus norvegicus*, *Spermophilus tridecemlineatus*, *Cavia porcellus*, *Oryctolagus cuniculus*, *Canis lupus, Felis catus*, *Erinaceus europaeus*, *Sorex araneus*, *Bos Taurus*, *Myotis lucifugus*, *Loxodonta Africana*, *Echinops telfairi*, *Dasypus novemcinctus*, *Monodelphis domestica, Ornithorhynchus anatinus*, *Gallus gallus*, *Xenopus tropicalis*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Danio rerio*, *Ciona intestinalis*, *Ciona savignyi*, *Aedes aegypti*, *Anopheles gambiae*, *Drosophila melanogaster*, *Caenorhabditis elegans,* and *Arabidopsis thaliana*. Fungi organisms: *Kluyveromyces lactis*, *Eremothecium gossypii*, *Candida glabrata*, *Debaryomyces hansenii*, *Saccharomyces cerevisiae*, *Pichia stipitis*, *Yarrowia lipolytica*, *Neurospora crassa , Gibberella zeae*, *Aspergillus fumigatus*, *Schizosaccharomyces pombe*, *Ustilago maydis*, *Filobasidiella neoformans*, and *Encephalitozoon cuniculi*. Protista organisms: *Plasmodium falciparum, Cryptosporidium parvum, Leishmania infantum, Trypanosoma brucei, Dictyostelium discoideum, and Giardia lamblia.* Mean *EPI* of each organism were considered to evaluate the variation in *EPI* of each group. The whiskers represent the standard error. Asterisk (*) are equal among each other and different from number sign (#). $P<0.001$ ANOVA one-way, Bonferroni post-hoc.



Supplementary Figure S6. Evolutionary Distance Average versus number of proteins and number of organisms. 5% of the KOGs present in STRING database were sorted. The evolutionary distance among all proteins of each KOG evaluated was calculated and the evolutionary distance average (Poisson Distance) was obtained. Poisson Distance was plotted against number of proteins (A) and number of organisms (B) of each KOG evaluated.



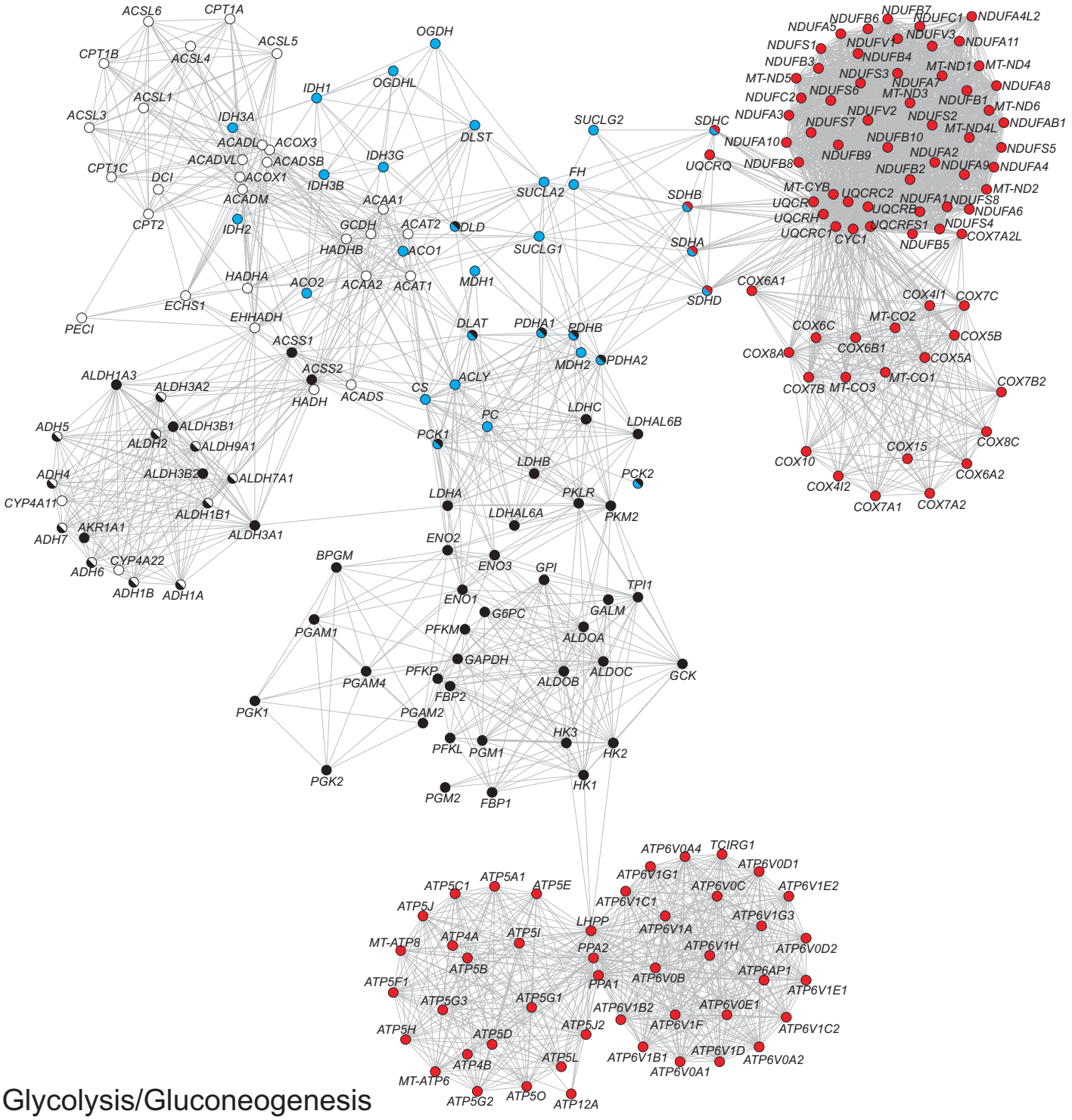
Supplementary Figure S7. Human and yeast ribosome gene networks. Ribosome protein-protein interaction networks of *H. sapiens* and *S cerevisiae* are shown. Nodes represent genes and the links represent protein-protein interaction of gene products.



- Glycolysis/Gluconeogenesis
- Fatty acid metabolism
- TCA cycle
- Oxidative phosphorylation

S. cerevisiae

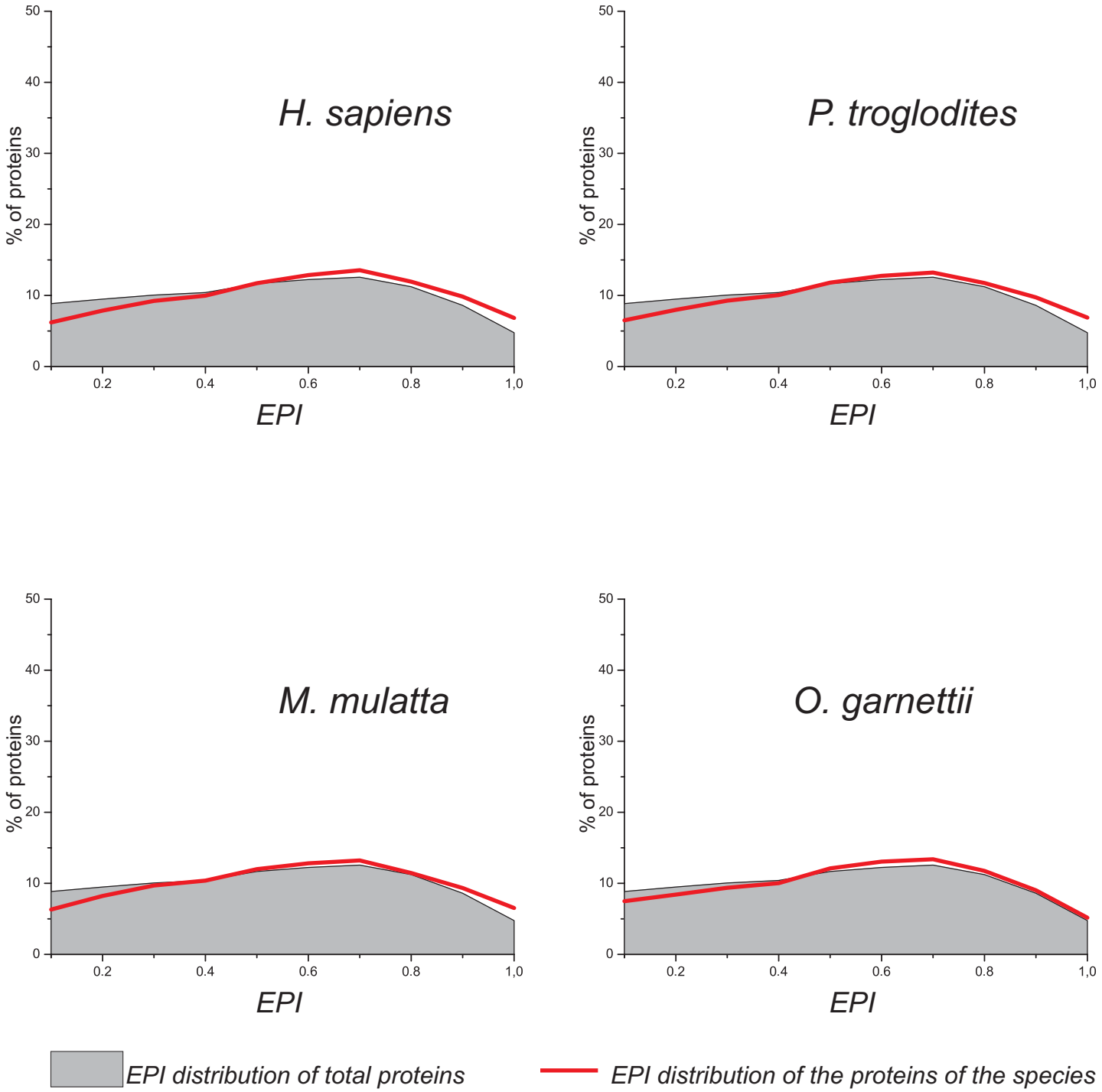
Supplementary Figure S8. *S. cerevisiae* energetic metabolism gene network. Energetic metabolism protein-protein interaction network of *S. cerevisiae* is shown. Nodes represent genes and links represent protein-protein interaction of gene products. Nodes were colored according to the pathways they belong. Nodes with more than one color belong to more than one pathway evaluated.



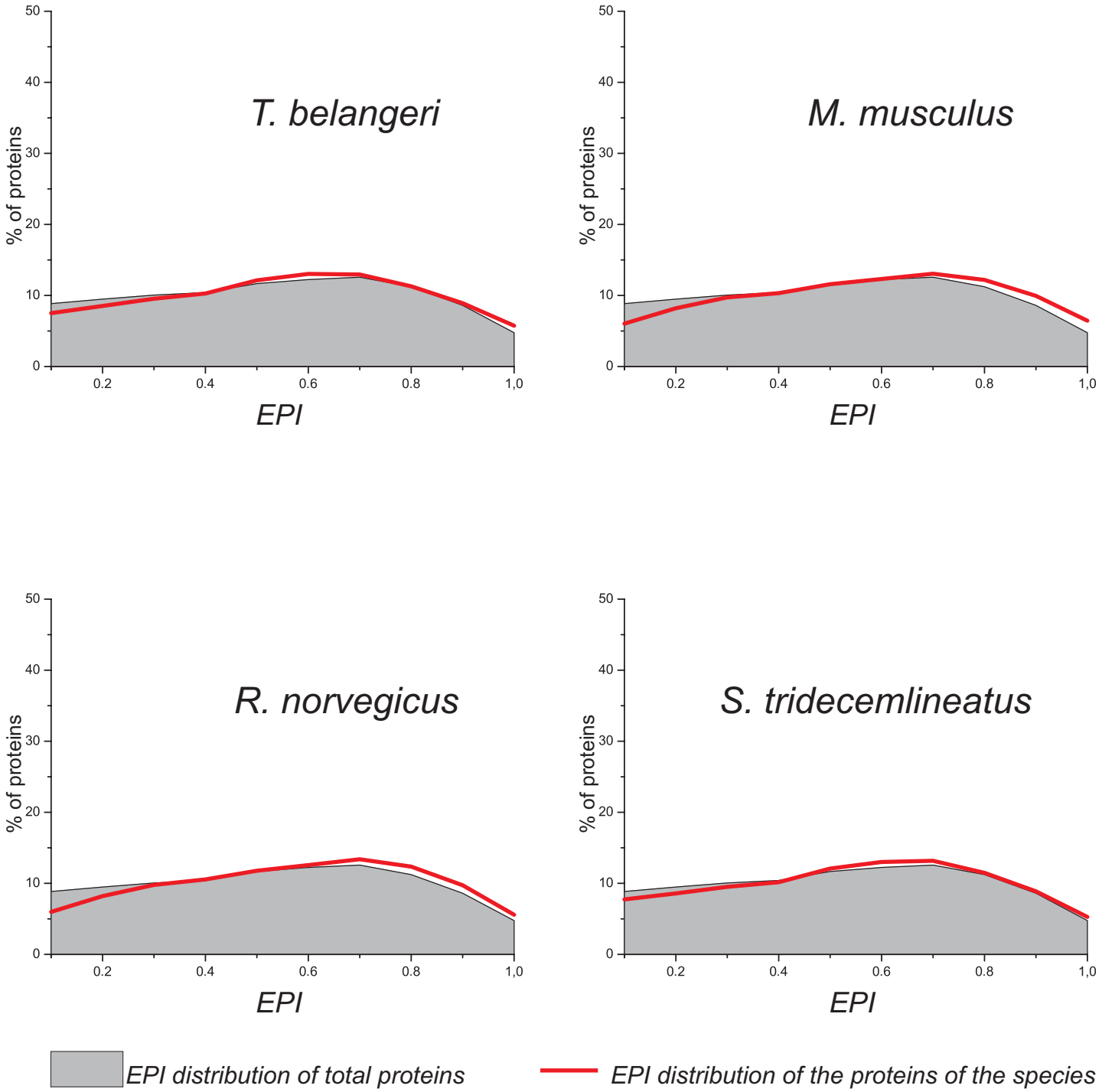
- Glycolysis/Gluconeogenesis
- Fatty acid metabolism
- TCA cycle
- Oxidative phosphorylation

H. sapiens

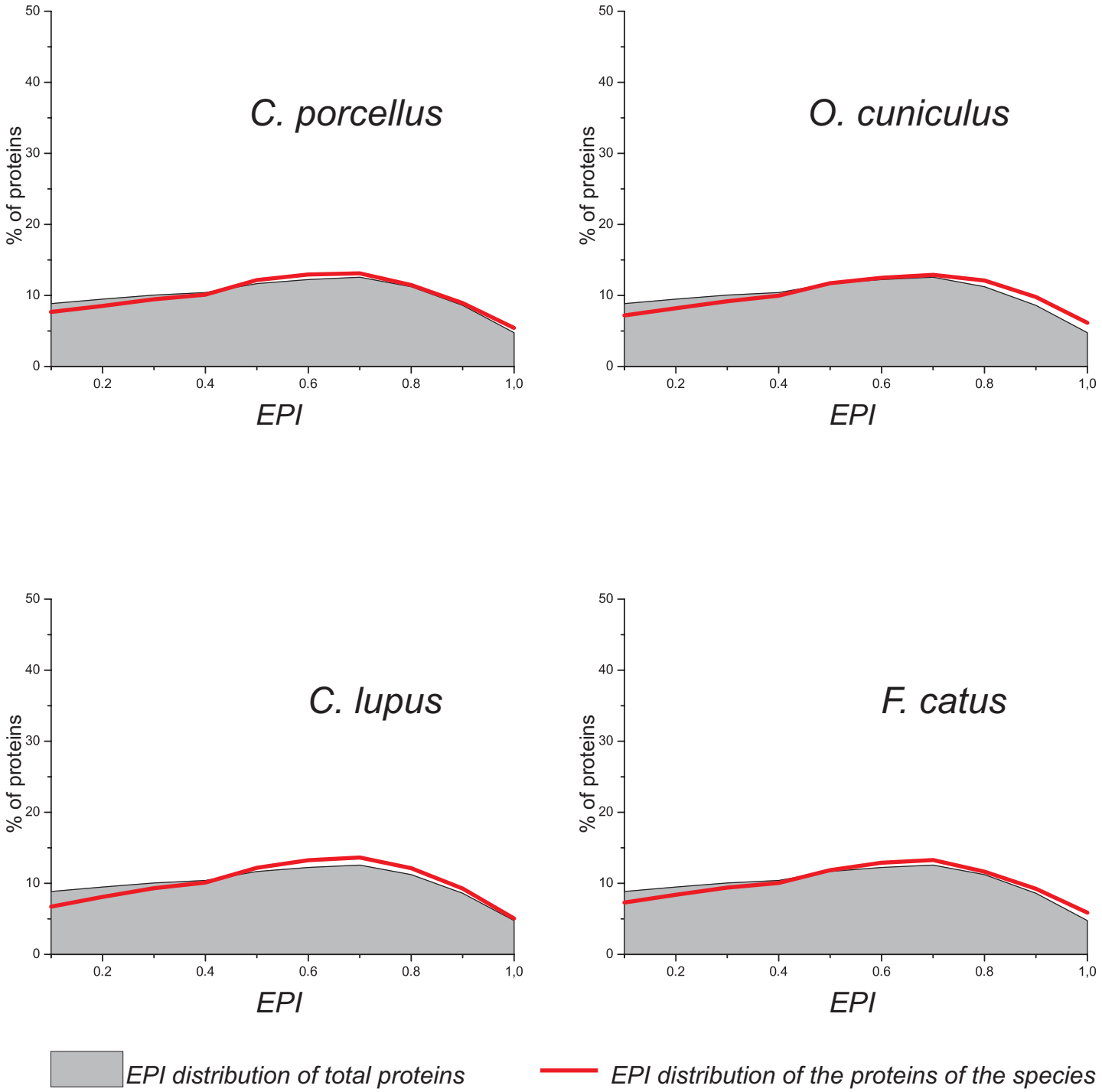
Supplementary Figure S9. *H. sapiens* energetic metabolism gene network. Energetic metabolism protein-protein interaction network of *H. sapiens* is shown. Nodes represent genes and links represent protein-protein interaction of gene products. Nodes were colored according to the pathways they belong. Nodes with more than one color belong to more than one pathway evaluated.



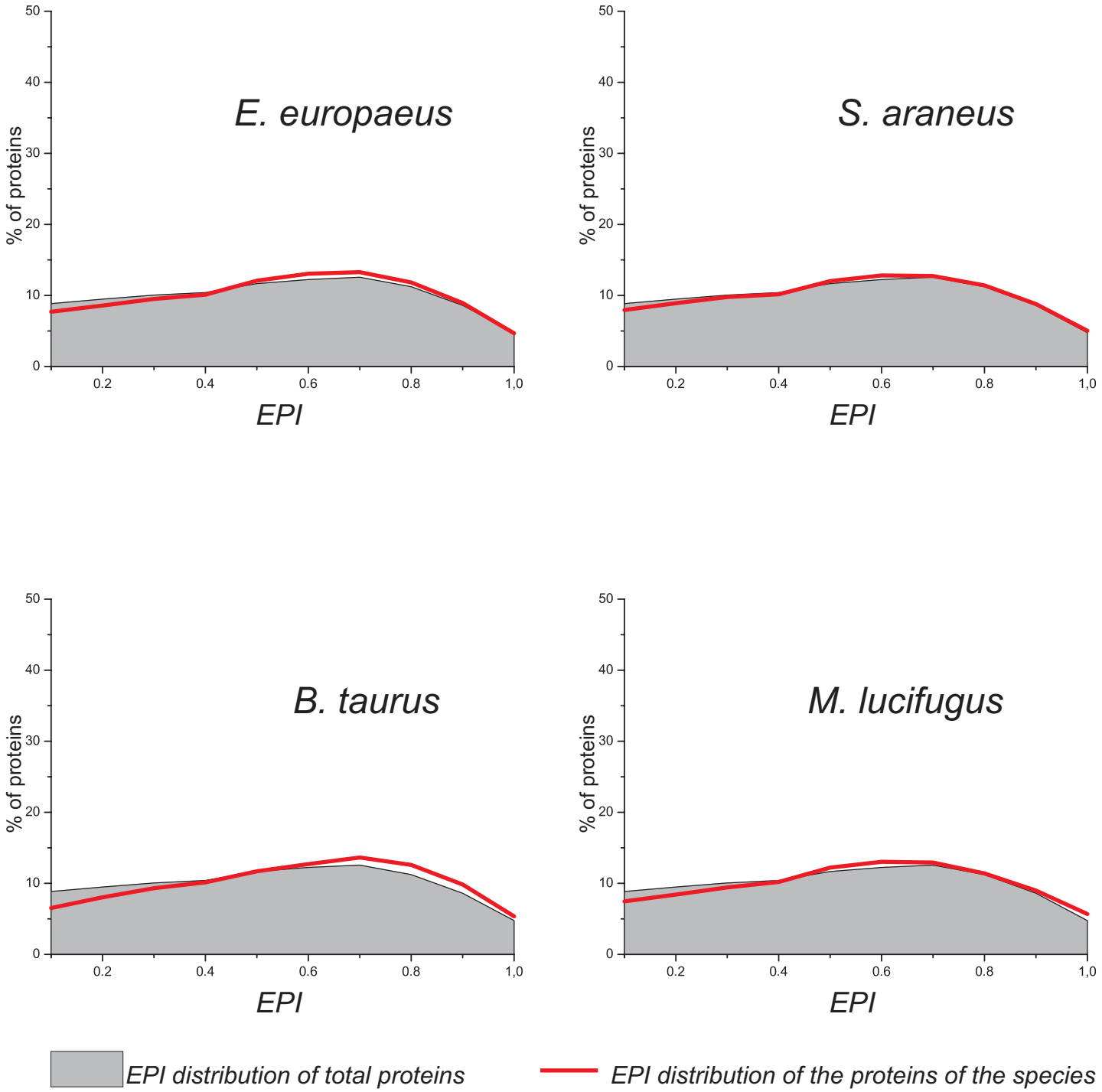
Supplementary Figure S10. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



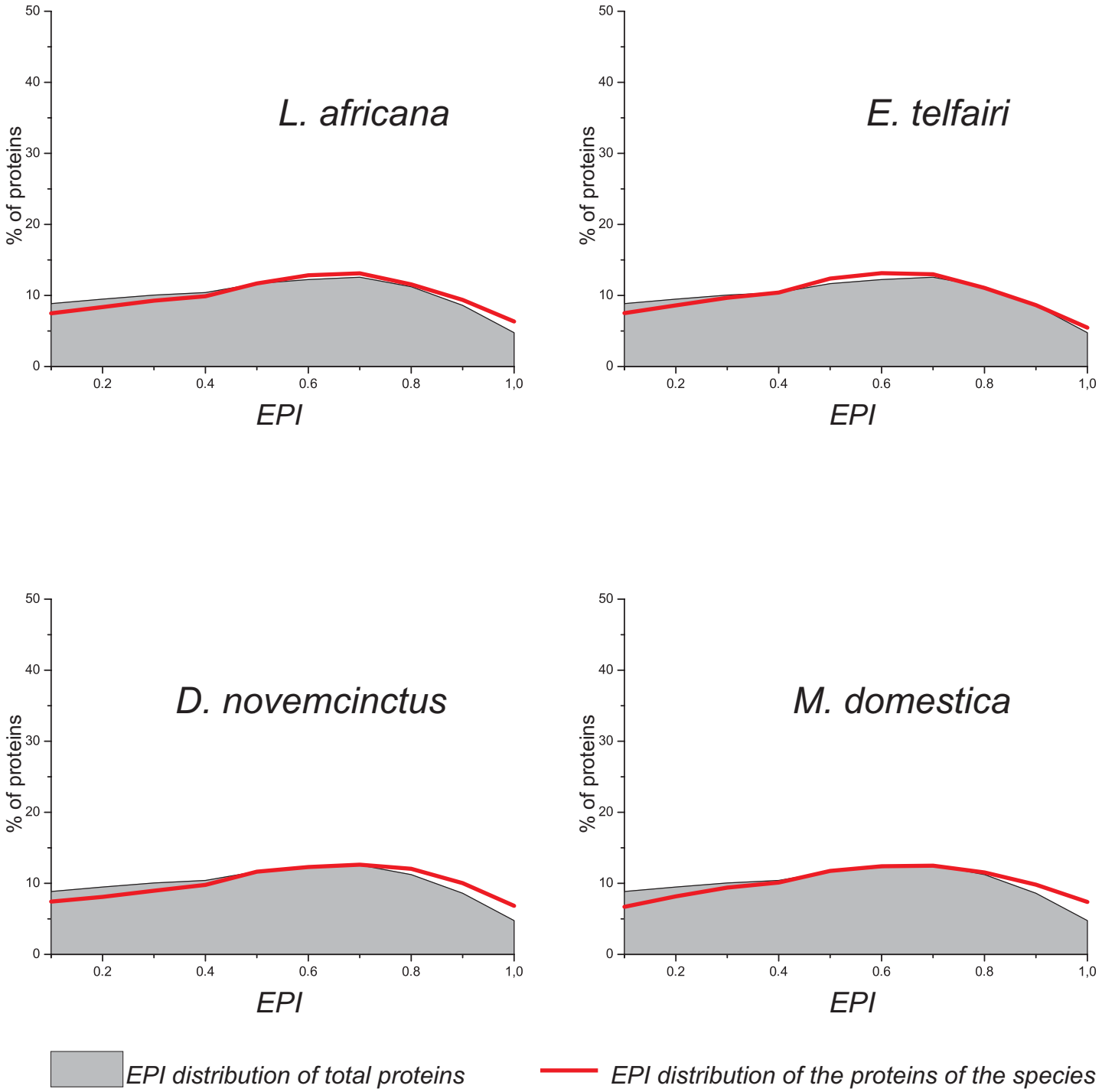
Supplementary Figure S11. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



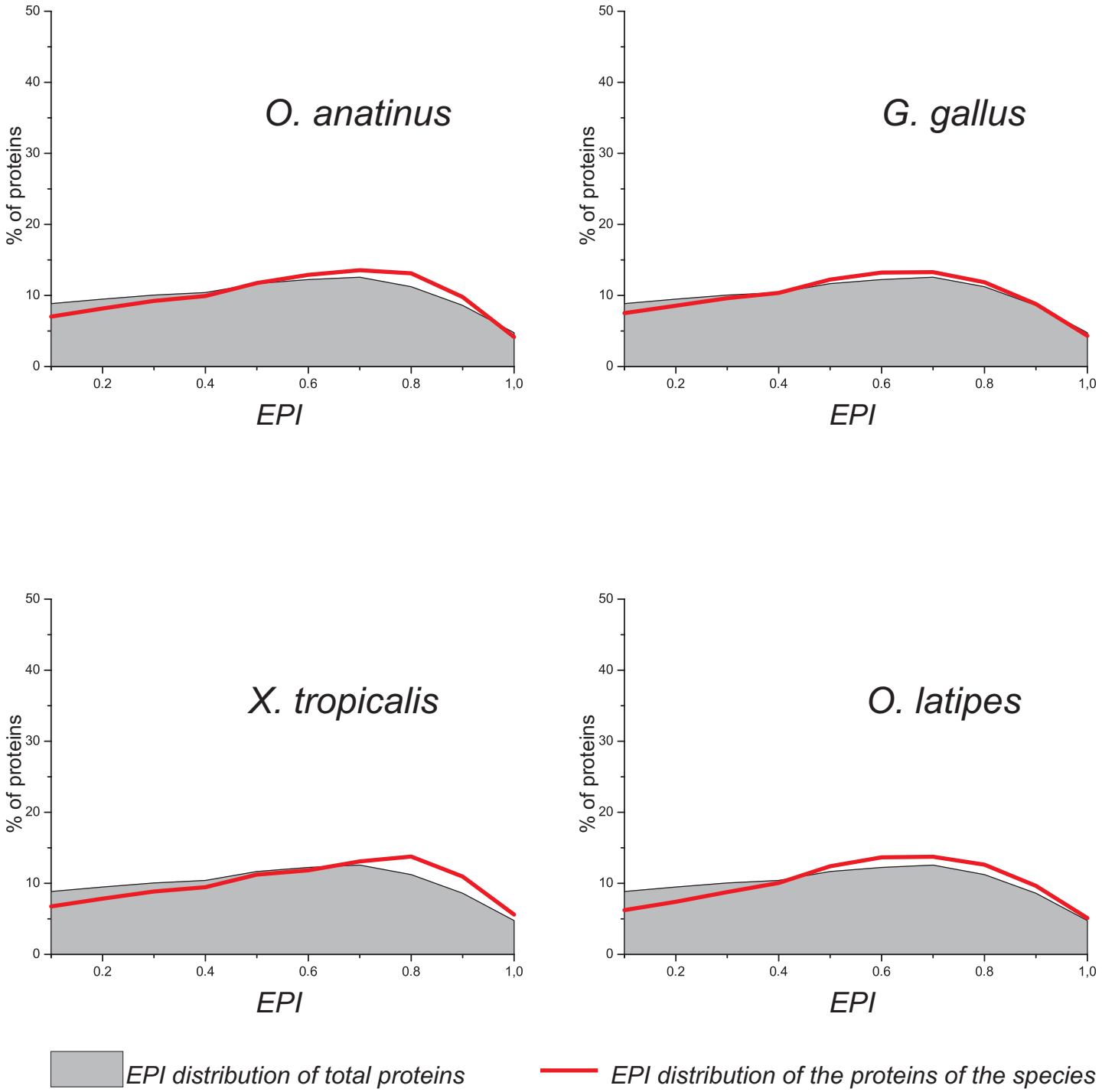
Supplementary Figure S12. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



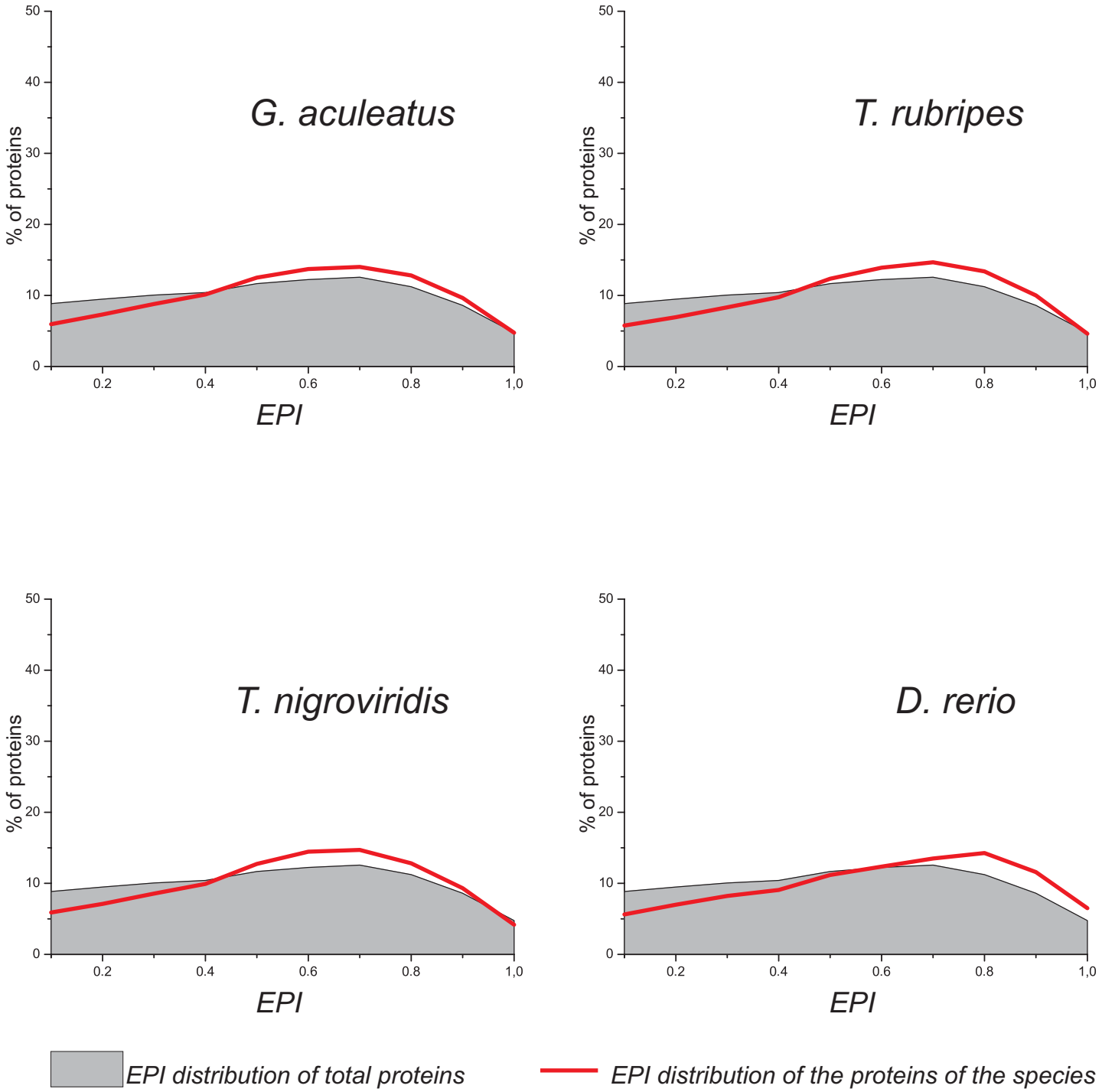
Supplementary Figure S13. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



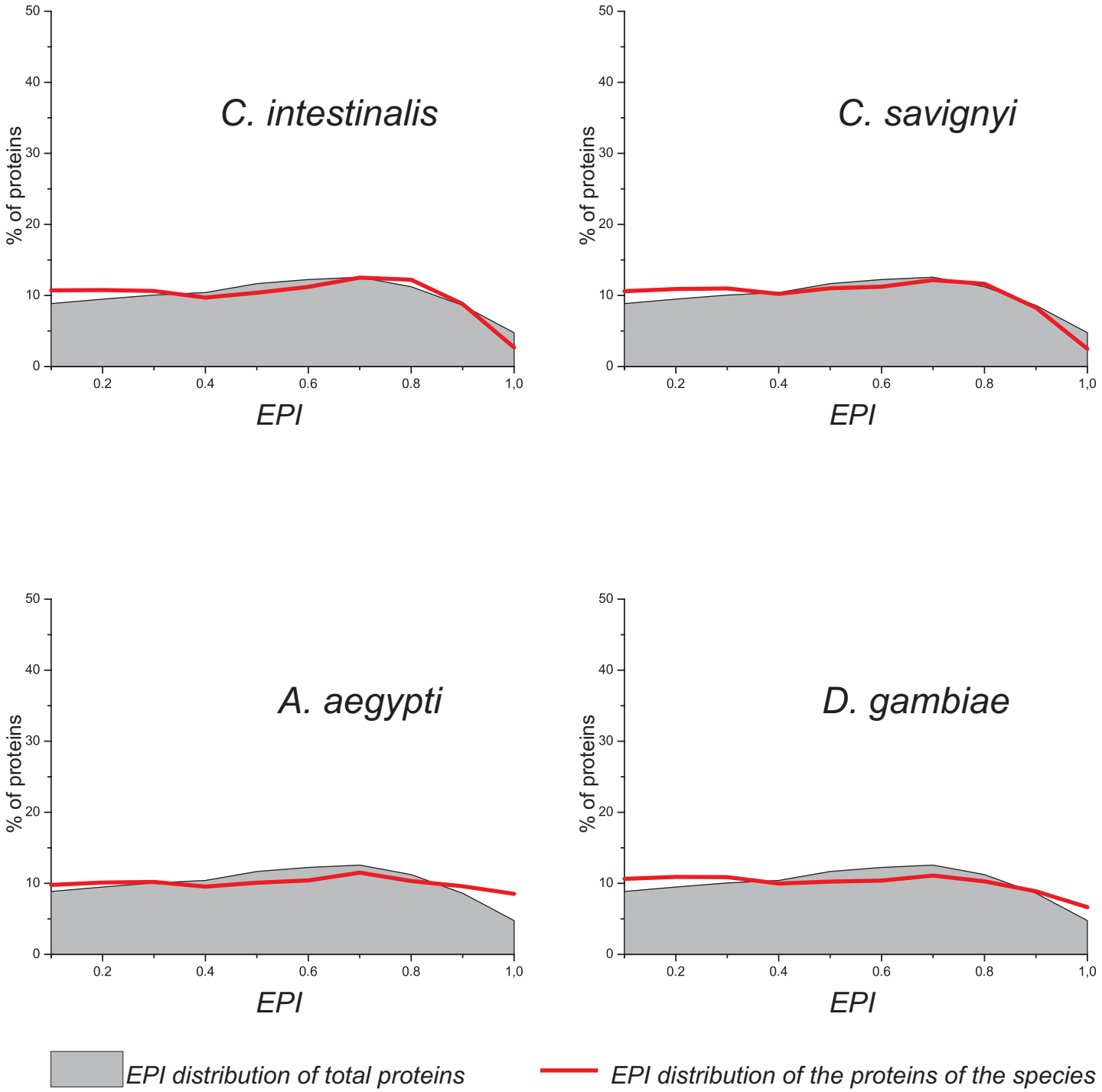
Supplementary Figure S14. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



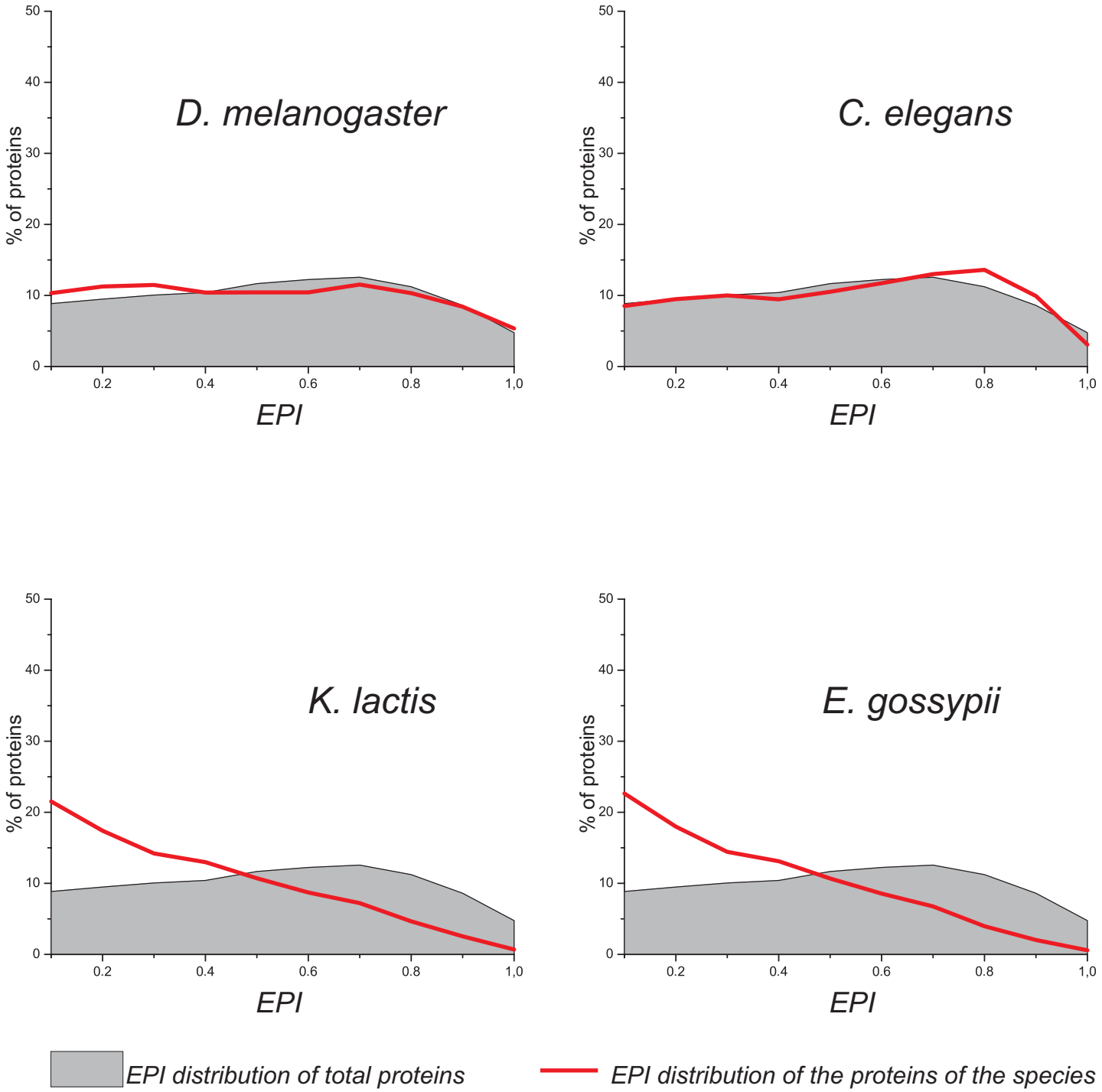
Supplementary Figure S15. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



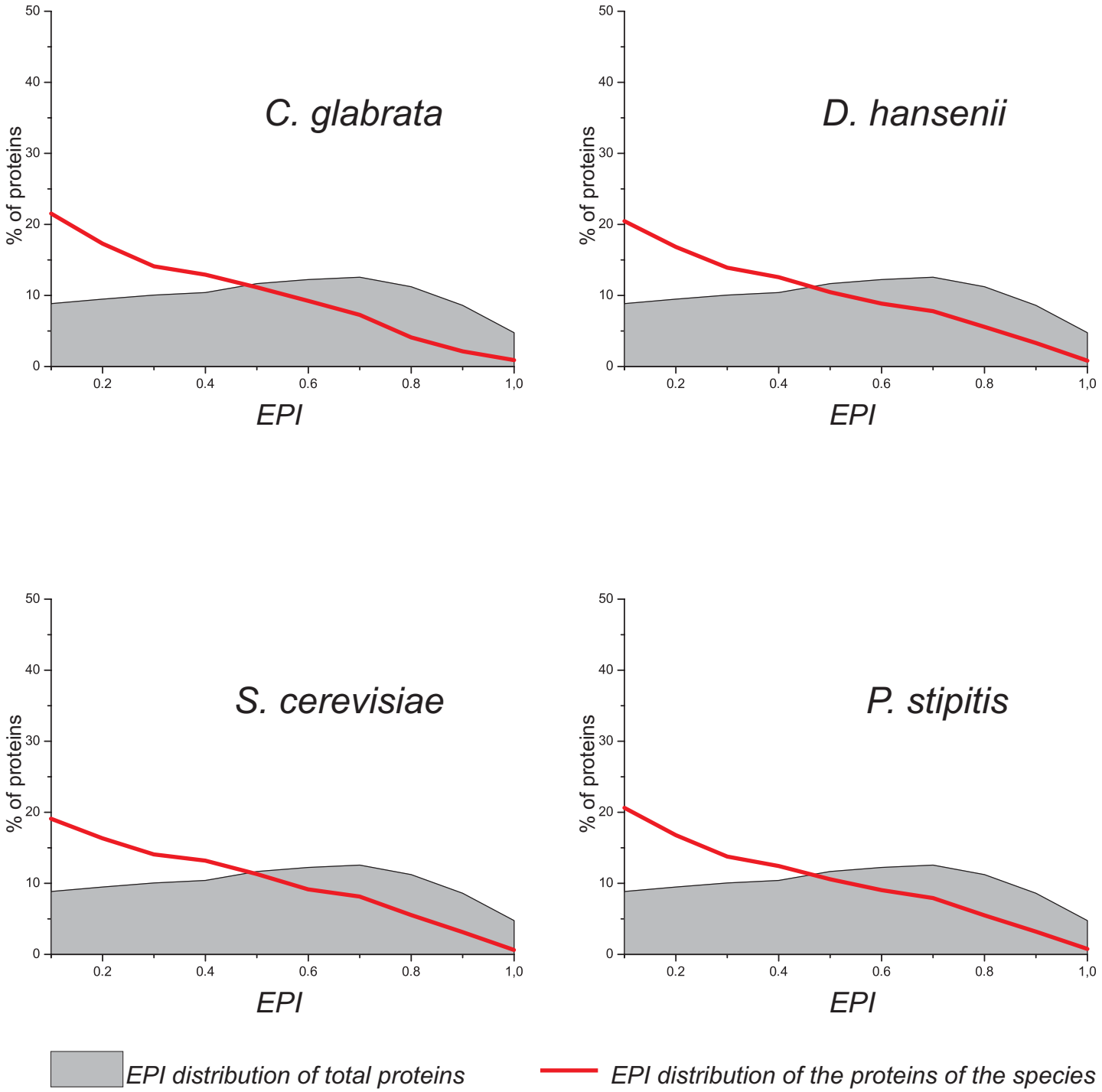
Supplementary Figure S16. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



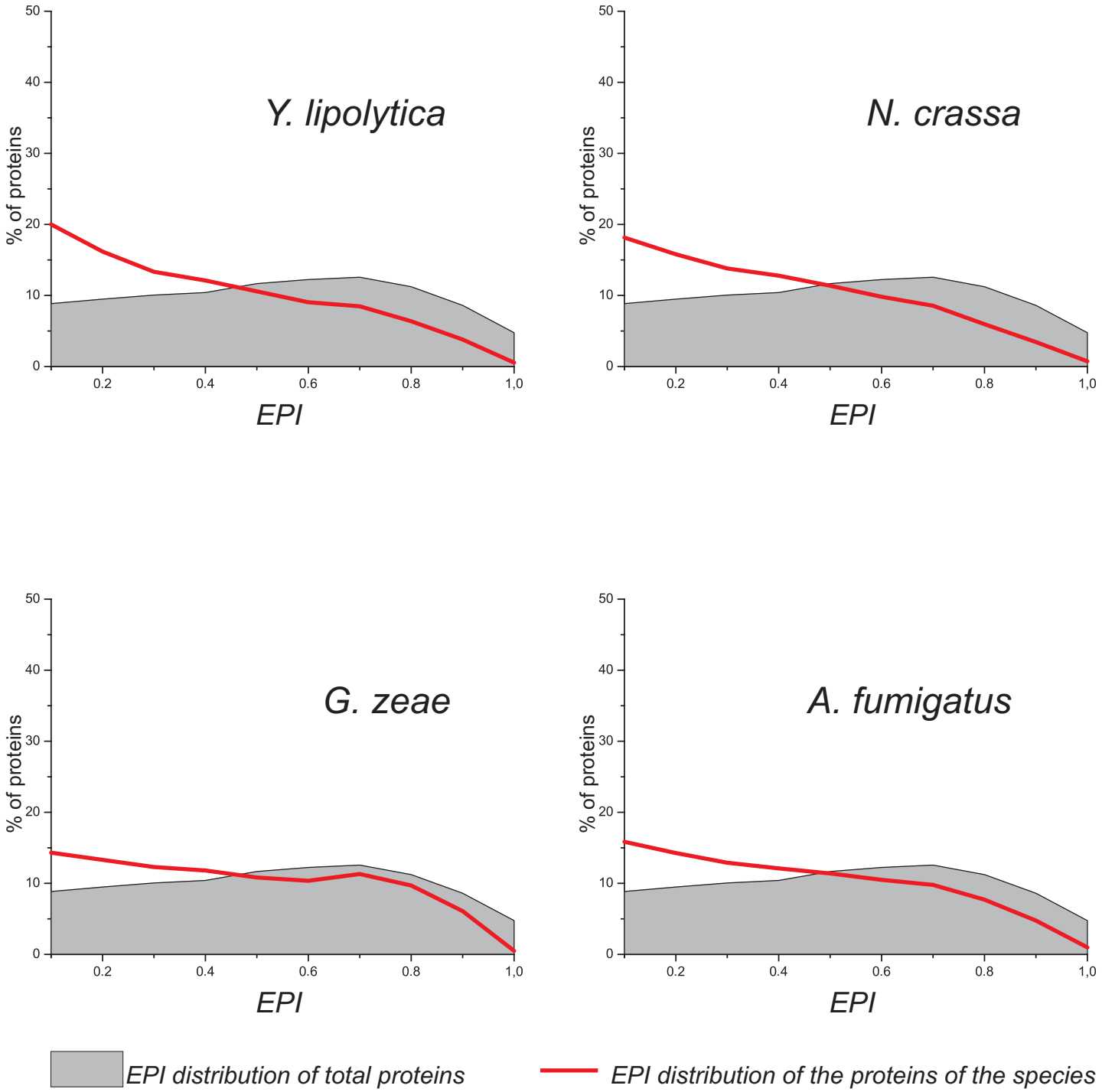
Supplementary Figure S17. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



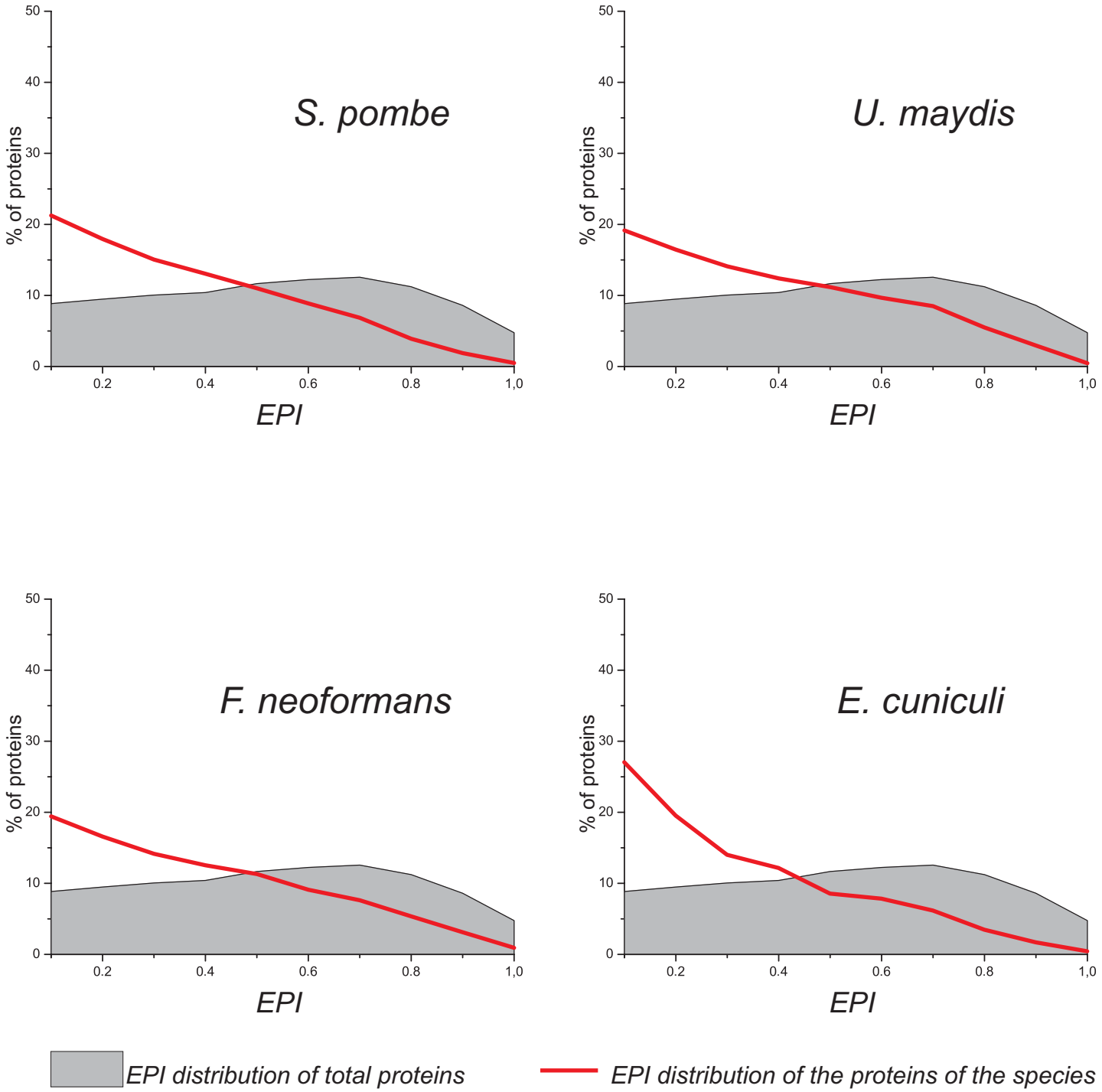
Supplementary Figure S18. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



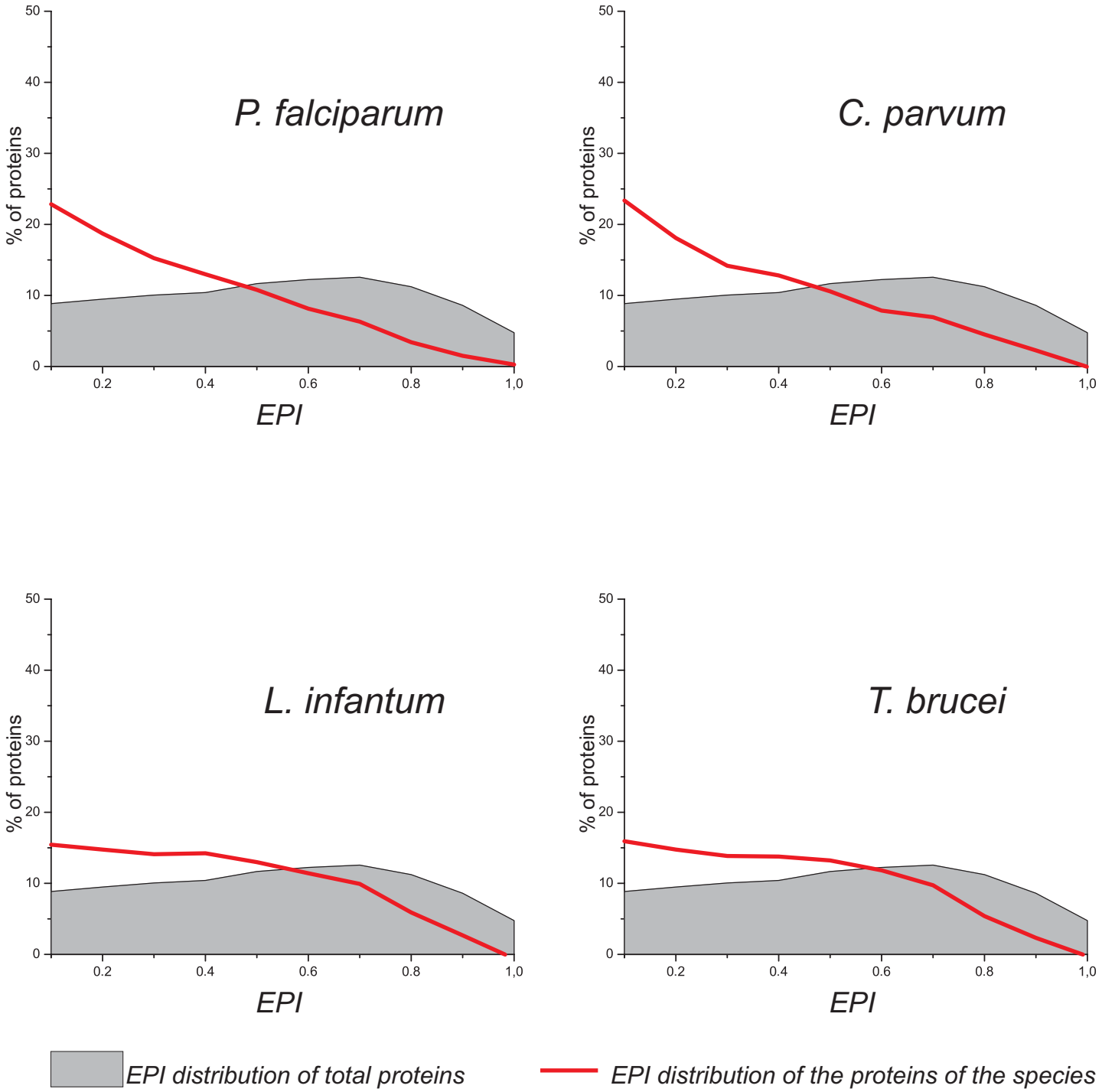
Supplementary Figure S19. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



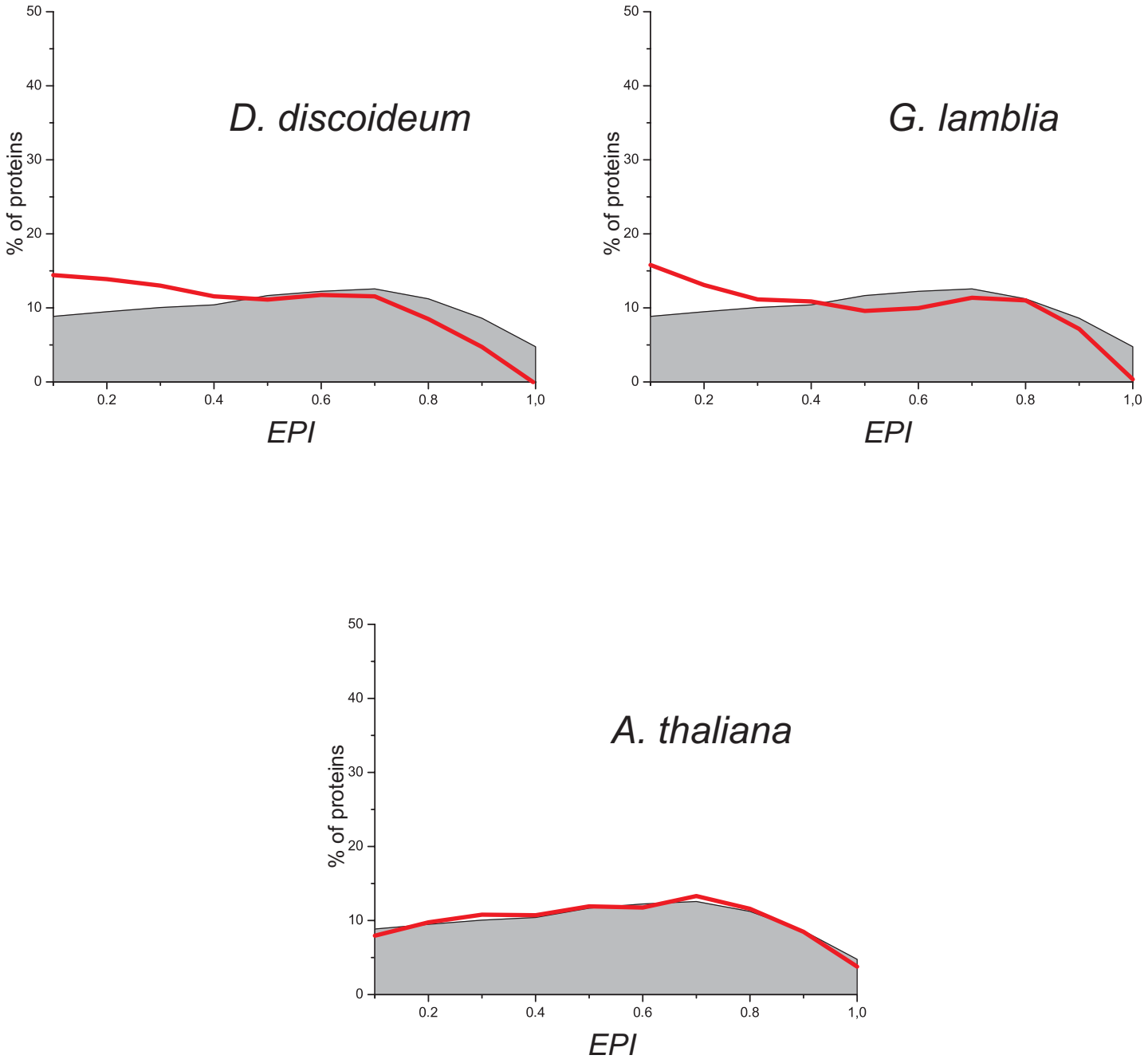
Supplementary Figure S20. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



Supplementary Figure S21. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



Supplementary Figure S22. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.



Supplementary Figure S23. *EPI* distribution of different species. Grey landscape represents the *EPI* distribution of all proteins present in KOG dataset and red line represents the *EPI* distribution of all proteins of each species.

2. References

1. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucl Acids Res* 2000, **28:** 27-30.

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

3. O'Brien KP, Remm M, Sonnhammer ELL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucl Acids Res* 2005, **33:** D476-D480.

4. Koonin EV, Makarova KS, Aravind L: **HORIZONTAL GENE TRANSFER IN PROKARYOTES: Quantification and Classification1.** *Annual Review of Microbiology* 2001, **55:** 709-742.

5. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J *et al*.: **STRING 8--a global view on proteins and their functional interactions in 630 organisms.** *Nucl Acids Res* 2009, **37:** D412-D416.

6. Dotsch A, Klawonn F, Jarek M, Scharfe M, Blocker H, Haussler S: **Evolutionary conservation of essential and highly expressed genes in Pseudomonas aeruginosa.** *BMC Genomics* 2010, **11:** 234.

7. He X, Zhang J: **Higher Duplicability of Less Important Genes in Yeast Genomes.** *Mol Biol Evol* 2006, **23:** 144-151.

8. Liao BY, Scott NM, Zhang J: **Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins.** *Mol Biol Evol* 2006, **23:** 2072-2080.

9. Liao BY, Zhang J: **Mouse duplicate genes are as essential as singletons.** *Trends in Genetics* 2007, **23:** 378-381.

10. Carroll SB: **Chance and necessity: the evolution of morphological complexity and diversity.** *Nature* 2001, **409:** 1102-1109.

11. Castro MA, Dalmolin RJ, Moreira JC, Mombach JC, de Almeida RM: **Evolutionary origins of human apoptosis and genome-stability gene networks.** *Nucleic Acids Res* 2008, **36:** 6269-6283.

12. Chen S, Zhang YE, Long M: **New Genes in Drosophila Quickly Become Essential.** *Science* 2010, **330:** 1682-1685.

13. Aravind L, Walker DR, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucl Acids Res* 1999, **27:** 1223-1242.

14. Igaki T, Kanda H, Yamamoto-Goto Y, Kanuka H, Kuranaga E, Aigaki T *et al*.: **Eiger, a TNF superfamily ligand that triggers the Drosophila JNK pathway.** *EMBO Journal* 2002, **21:** 3009-3018.

Evolution signatures in genome networks – Material suplementar

Supplementary Materials for Evolution signatures in genome networks

Ricardo M. Ferreira*[2], José Luiz Rybarczyk-Filho*[2], Rodrigo J. S. Dalmolin*[3], Mauro A. A. Castro[1,2], José C. F. Moreira[3], Leonardo G. Brunnet[2] & Rita M. C. de Almeida[1,2]

National Institute of Science and Technology for Complex Systems[1], Instituto de Física[2], and Departamento de Bioquímica[3], Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, 91051-970 C.P. 15051, Porto Alegre, Brazil.

***These authors contributed equally to this paper**

Correspondence to: Rita M. C. de Almeida (rita@if.ufrgs.br).

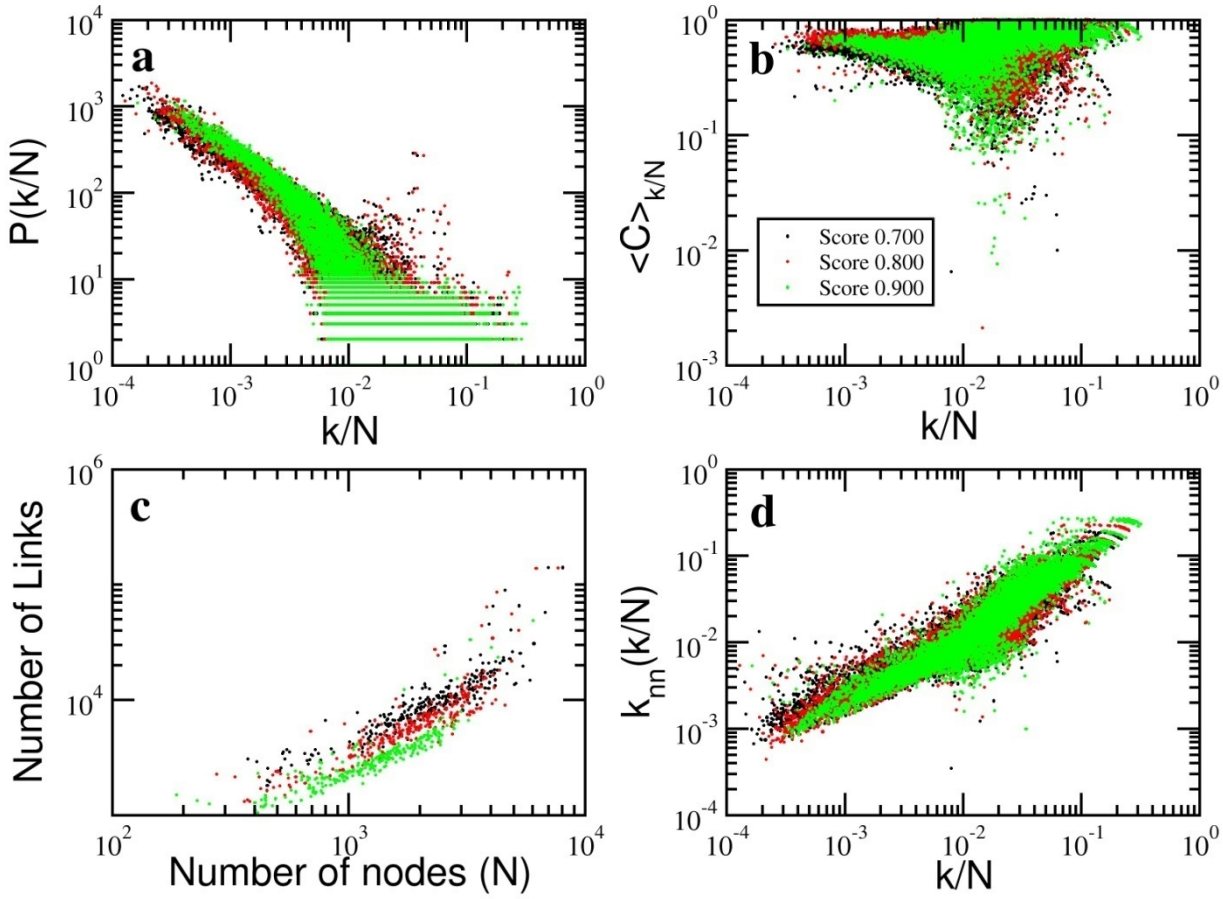


Figure S1. Topological measures for all core organisms in STRING database with confidence scores of 0.700, 0.800, and 0.900 (black, red and green dots), with degree rescaled by number of nodes *N*. Figure (**a**) shows degree distribution, (**b**) clustering coefficient, (**c**) number of links per number of nodes, and (**d**) mean average degree of nearest neighbors. In these figures we can see that the properties discussed in the main text are not clearly evinced.

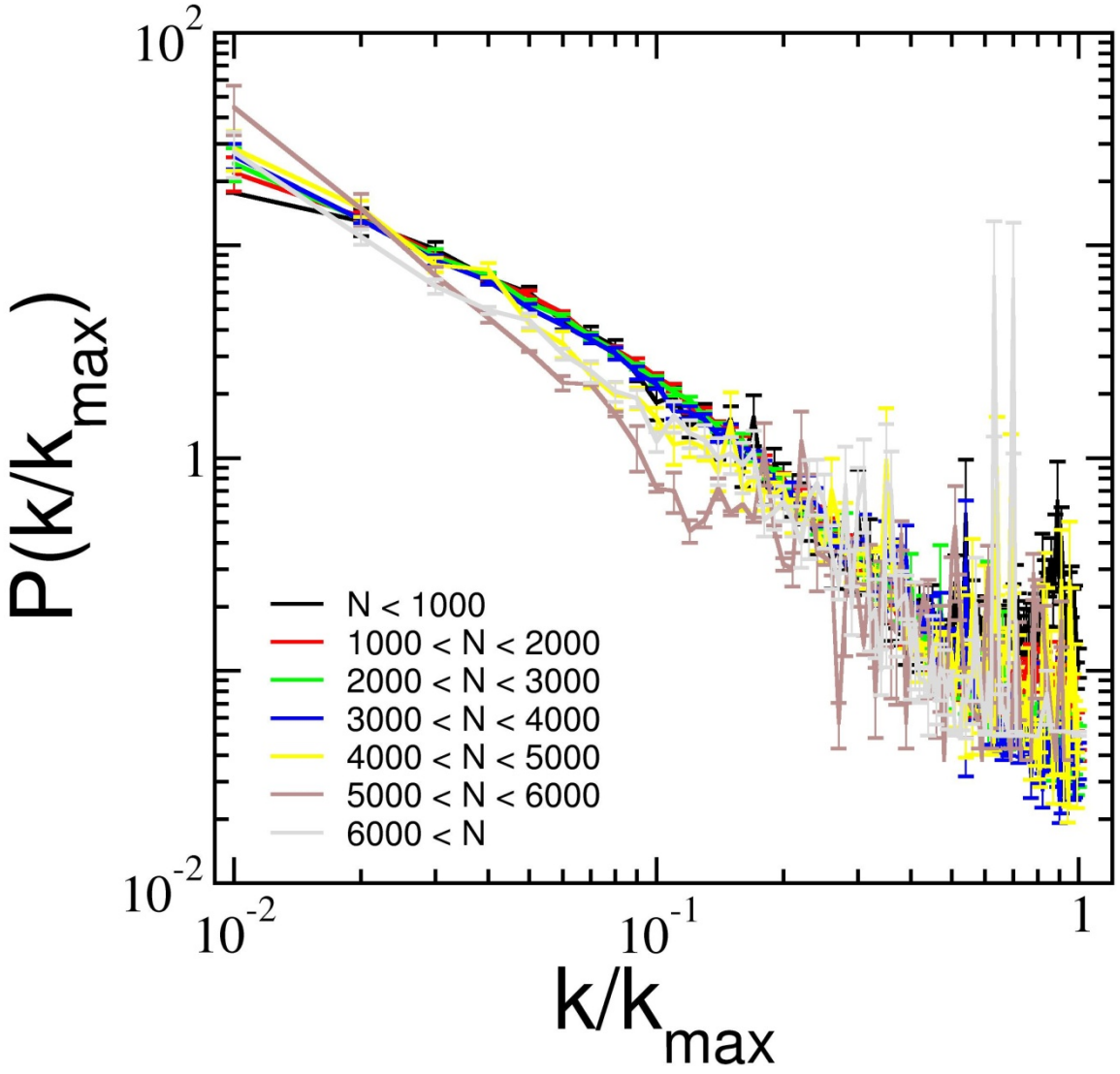


Figure S2. Degree distribution of protein-protein association matrices relative to core organisms for STRING confidence score 0.800, averaged over intervals of $\frac{k}{k_{\max}} = 0.01$. Each line corresponds to a network in a different range of number of nodes N, as described in the legend.

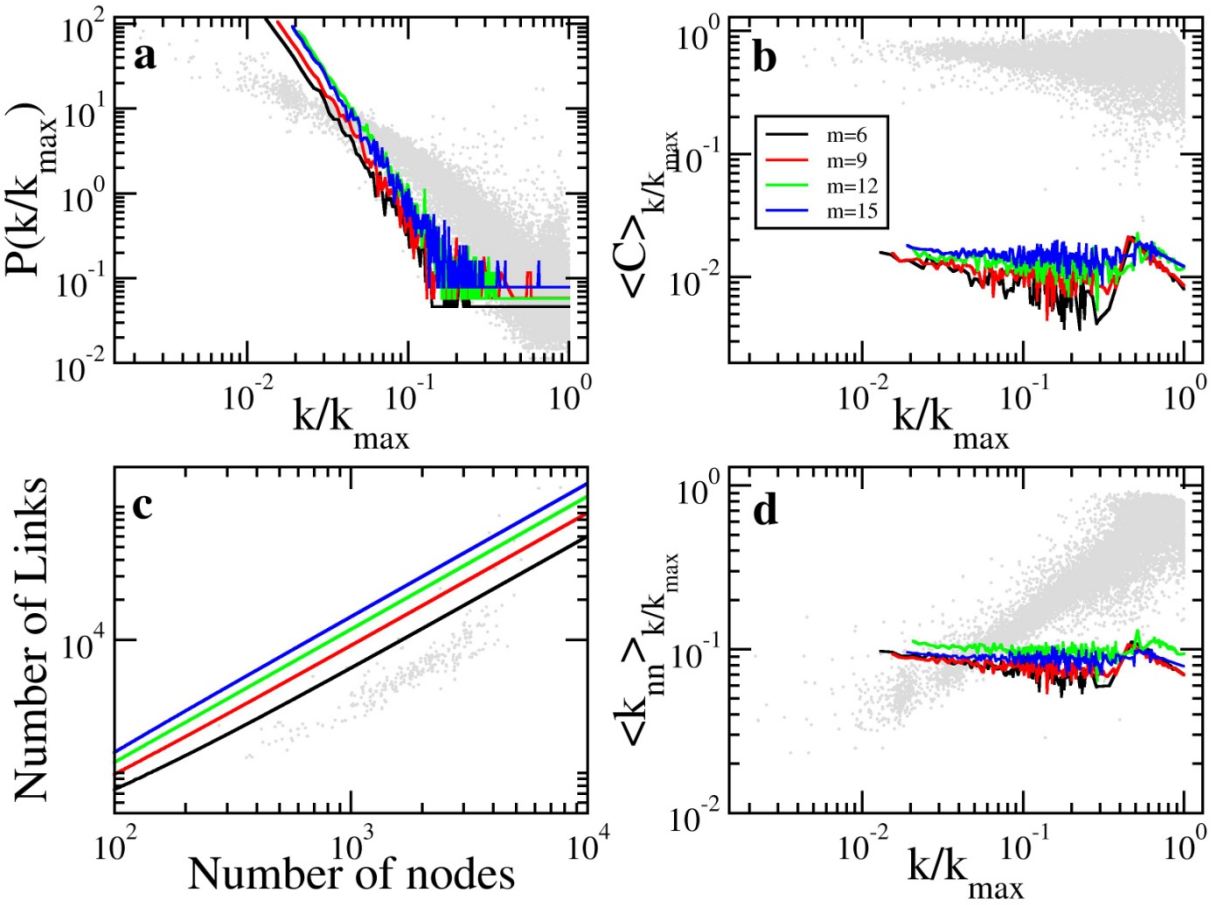


Figure S3. Four networks obtained using Barabási-Albert model with different values of parameter *m*, which determines the number of links of each new node. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. In Figure (**a**) we can see that the degree distribution follows a power-law and does not correctly represent the degree distribution of the organisms networks. Figure (**b**) shows that clustering coefficient of the simulated networks is lower than the experimental data. Figure (**c**) presents the evolution of number of links with number of nodes. Figure (**d**) shows that the average degree of the neighbors of a node is independent of the node degree for networks built using Barabási-Albert model, what deviates from the behavior presented by the organisms networks that are highly assortative.

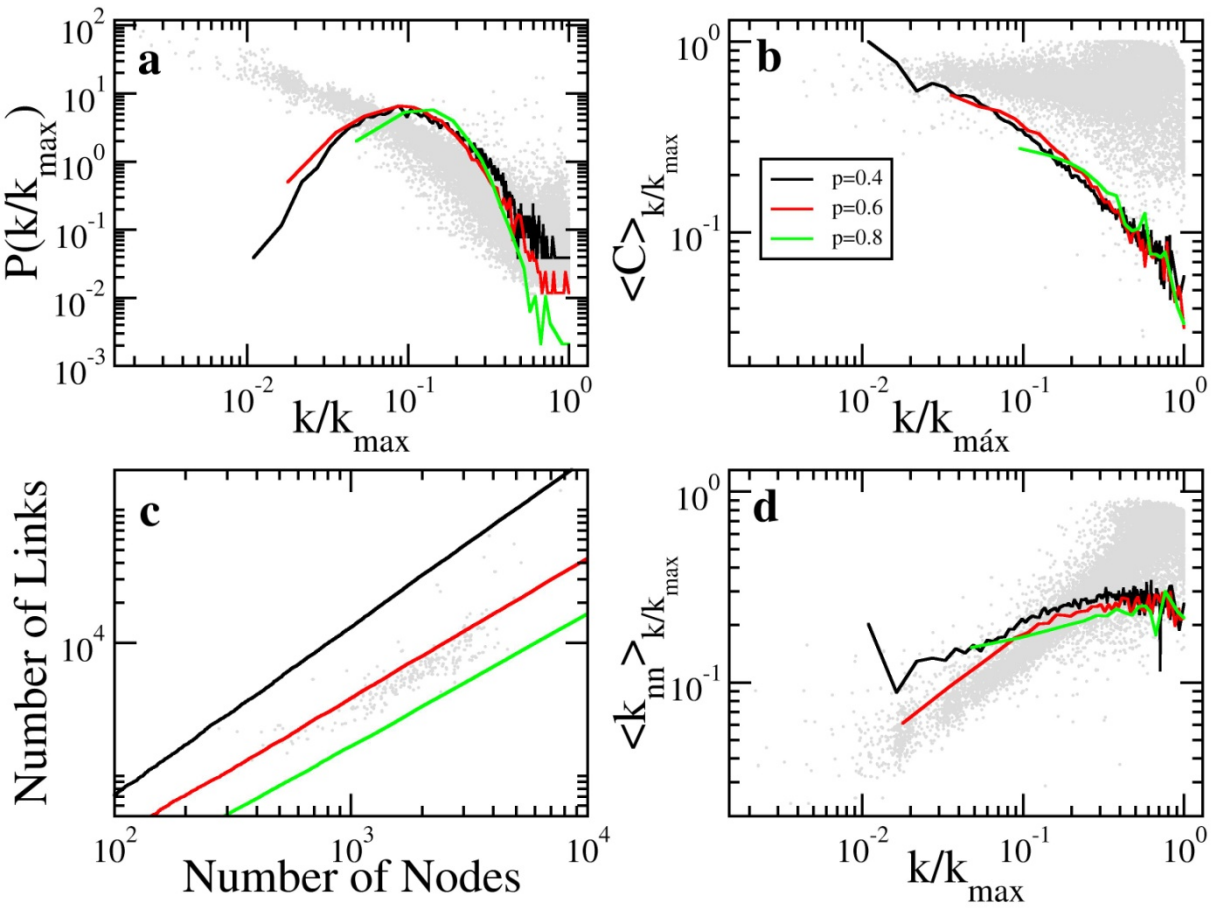


Figure S4. Three networks obtained using Duplication-Divergence model with different values of parameter *p*, which determines the mutation probability. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. In Figure (**a**) we can see that for higher values of *p* the network approaches a power-law, but as we can see in Figure (**c**), the number of links fall below those found for the organisms. Figure (**b**) shows that the clustering decreases with degree. In Figure (**d**) we have the average degree of nearest neighbors, which increases with degree, showing that the Duplication Divergence model builds networks with the same assortativeness of the organisms.

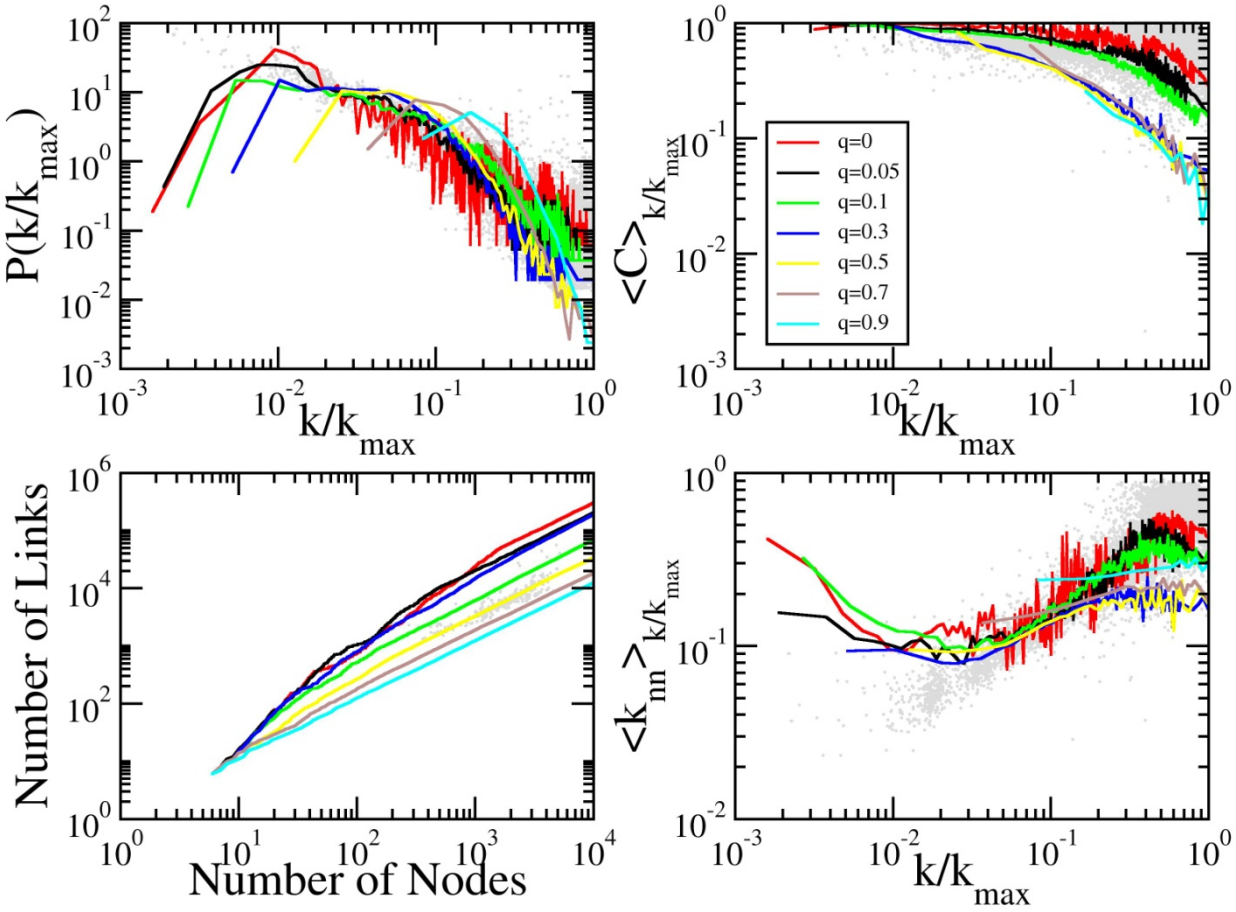


Figure S5. Five networks obtained using the Duplication-Acquisition model with different values of parameter *q*, which determines the fraction of nodes acquired by duplication, maintaining constant *r*, the mutation probability. The grey dots represent the networks for all 268 core organisms, with confidence score 0.800. We can see that, as the number of acquired nodes increases, the network approaches a Barabási-Albert one, as we can see in Figures (**a**), (**b**), and (**d**). Namely, the network loses the high probability of finding high degree nodes in the degree distribution (Figure (**a**)), the clustering coefficient decreases (Figure (**b**)), and the network loses its assortativity (Figure (**d**)). Figure (**c**) shows that the number of links also decreases, falling below the value presented by organisms.

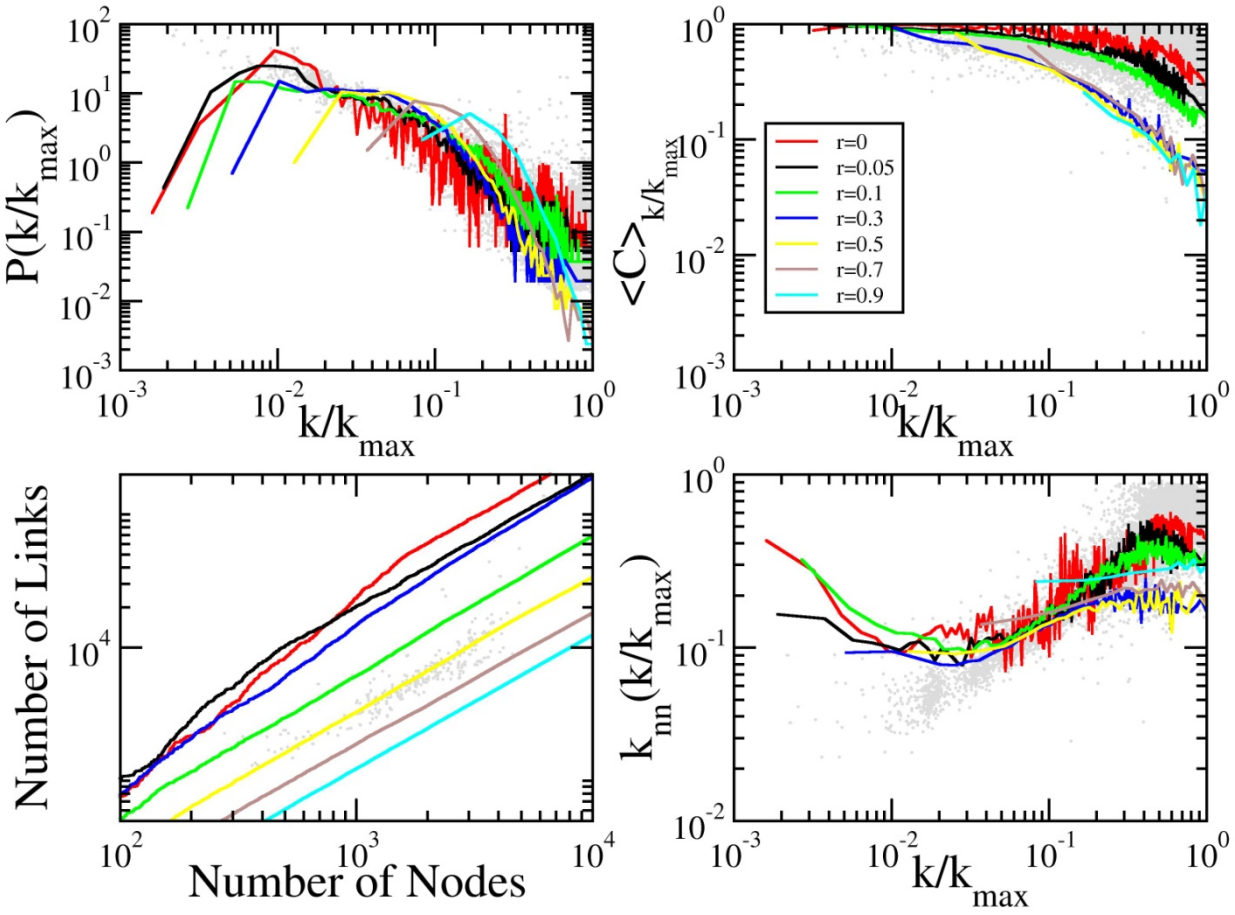


Figure S6. Seven networks obtained using the Duplication-Acquisition model with different values of parameter *r*, which determines the mutation probability, maintaining constant *r*, the fraction of nodes acquired by duplication. The grey dots represent networks for all 268 core organisms, with confidence score 0.800. We can see that, as mutation probability increases, the network approaches the ones obtained using the Duplication-Divergence model. In Figure (**a**) the degree distribution approaches a power-law, and in (**b**) the clustering coefficient decreases. Figure (**c**) shows that the number of links decreases, and in Figure (**d**) we can see the mean nearest degree distribution for the networks.

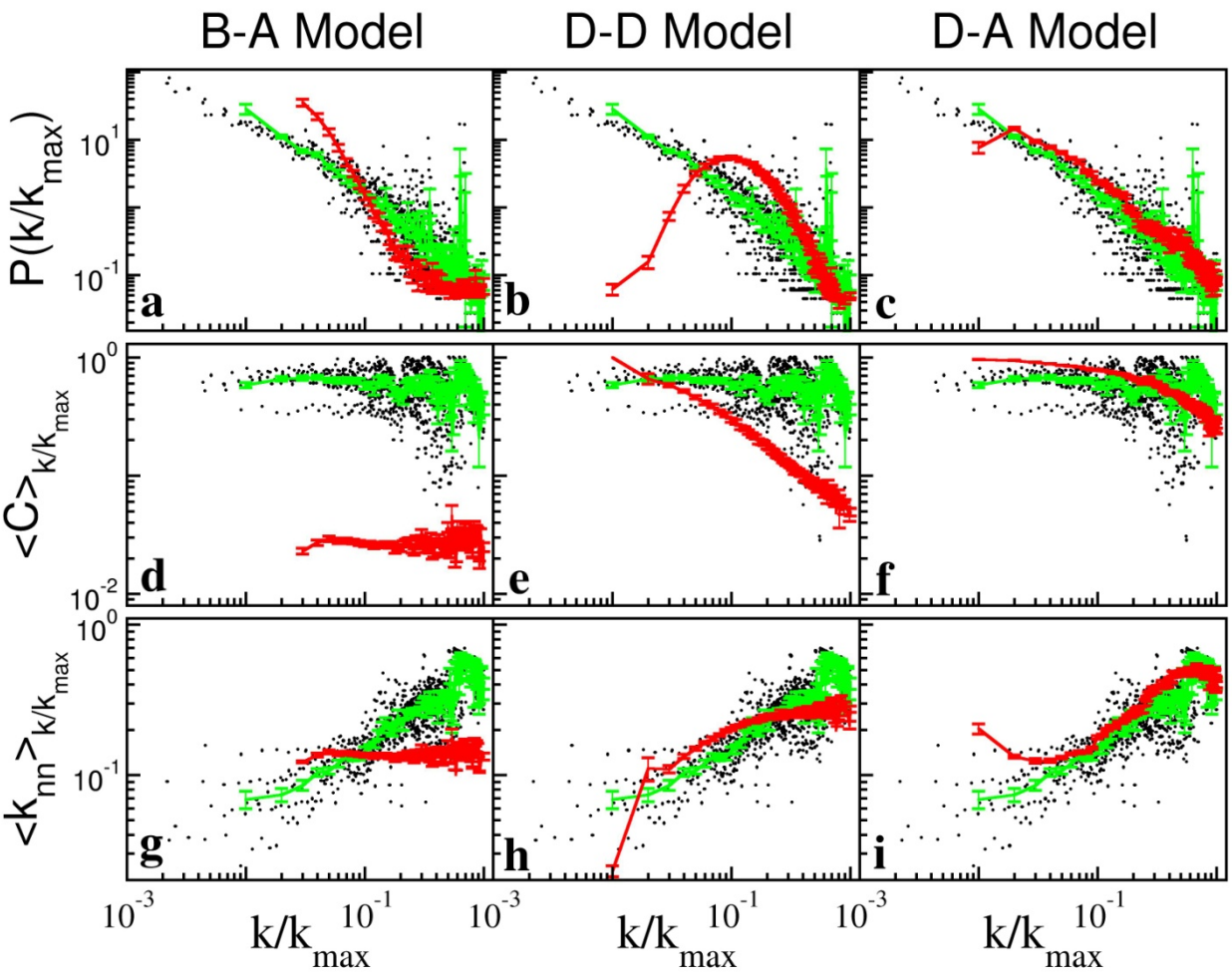


Figure S7. Comparison of topological measures for simulated networks. The black dots represent the superposed networks for six organisms from STRING database with confidence score 0.800 (*Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae,* and *Escherichia coli)*, the red lines are averages of these networks taken in intervals of $^{k}/_{k_{\max}} = 0.01$, and the green lines are weighted averages of simulated networks. Upper, central, and lower rows show, respectively, degree distribution, clustering coefficient, and nearest neighbor mean degree. Each column refers to a simulated model: Barabási-Albert on the left, duplication-divergence on the center and duplication-acquisition on the right.

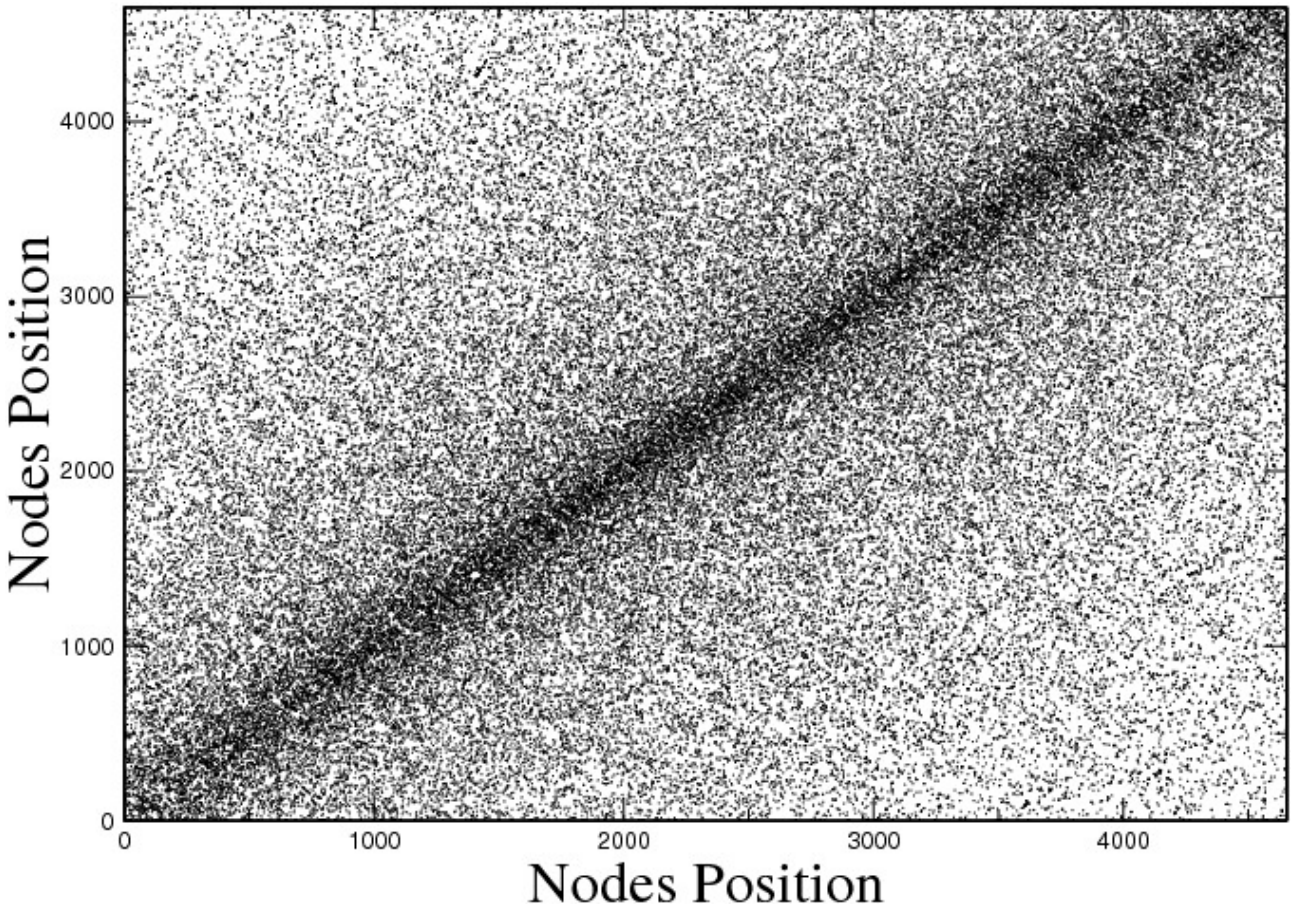


Figure S8. Association matrix for a Erdös-Rényi network with 4665 nodes and 94830 links, ordered using $\alpha=1$.

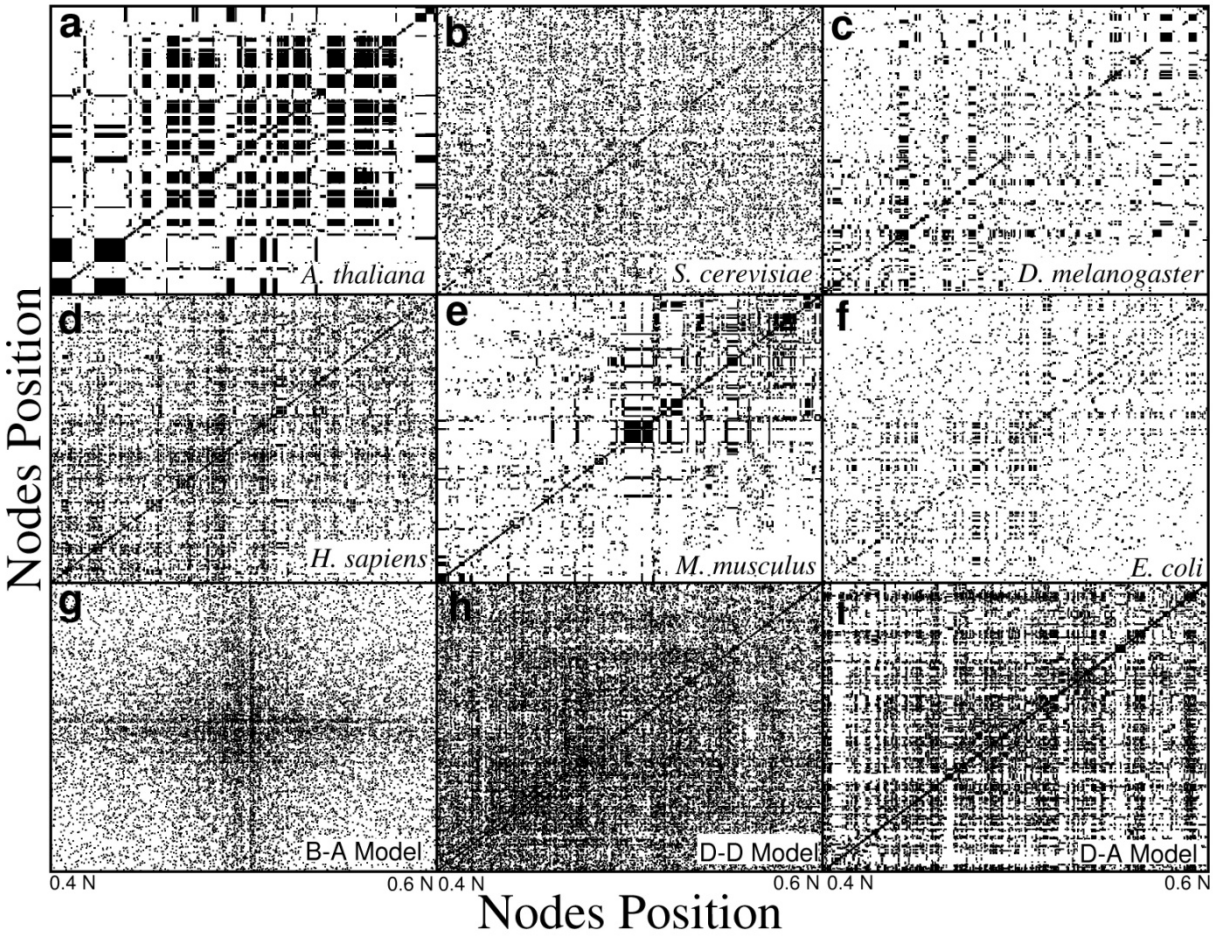


Figure S9. Zoom at the central part of association matrices in Fig.3, from 0.4N to 0.6 N, for **(a)** *Homo sapiens*, **(b)** *Mus musculus*, **(c)** *Arabidopsis thaliana*, **(d)** *Drosophila melanogaster*, **(e)** *Saccharomyces cerevisiae*, **(f)** *Escherichia coli*, **(g)** Barabási-Albert model, **(h)** duplication-divergence model and **(i)** duplication-acquisition model, ordered using α=8.

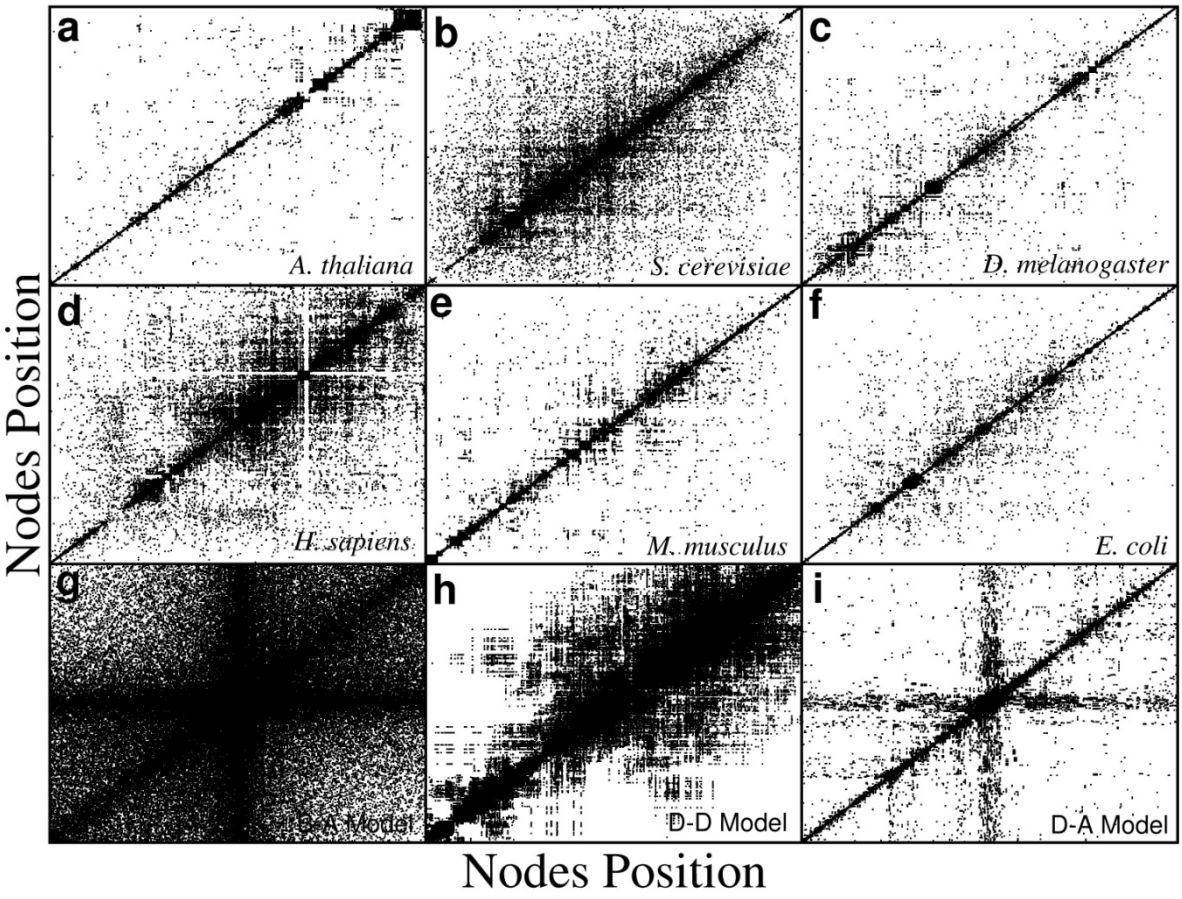


Figure S10. Association matrices for **(a)** *Homo sapiens*, **(b)** *Mus musculus*, **(c)** *Arabidopsis thaliana*, **(d)** *Drosophila melanogaster*, **(e)** *Saccharomyces cerevisiae*, **(f)** *Escherichia coli*, **(g)** Barabási-Albert model, **(h)** duplication-divergence model and **(i)** duplication-acquisition model, ordered using α=1.