A nested phylogenetic reconstruction approach provides scalable resolution in the eukaryotic Tree Of Life

Jaime Huerta-Cepas[1,2], Marina Marcet-Houben[1,2], and Toni Gabaldón[1,2,*]

1- Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG) Doctor Aiguader, 88. 08003 Barcelona (Spain)

2- Universitat Pompeu Fabra (UPF). 08003 Barcelona (Spain)

*Corresponding Author. e-mail: tgabaldon@crg.eu .
Telephone: +34 933160281.
Fax: +34 93 3969983

Abstract

Assembling the Tree Of Life (TOL) faces the pressing challenge of incorporating a rapidly growing number of sequenced genomes. This problem is exacerbated by the fact that different sets of genes are informative at different evolutionary scales. Here, we present a novel phylogenetic approach (Nested Phylogenetic Reconstruction) in which each tree node is optimized based on the genes shared at that taxonomic level. We apply such procedure to reconstruct a 216-species eukaryotic TOL and compare it with a standard concatenation-based approach. The resulting topology is highly accurate, and reveals general trends such as the relationship between branch lengths and genome content in eukaryotes. The approach lends itself to continuous update, and we show this by adding 29 and 173 newly-sequenced species in two consecutive steps. The proposed approach, which has been implemented in a fully-automated pipeline, enables the reconstruction and continuous update of highly-resolved phylogenies of sequenced organisms.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

Introduction

The advent of genomics carried the promise of using the full genetic complement of species to unravel their evolutionary relationships. Efforts towards this aim have mainly focused on the combined analysis of multiple genes (Delsuc et al., 2005), by, for instance, concatenating their alignments. This so-called gene concatenation –or super-matrix- approach has the advantage over alternative widespread strategies (e.g. super-trees (Bininda-Emonds, 2004)) of using directly the information contained in the substitution patterns of homologous residues (Delsuc et al., 2005), and of providing branch length estimates. A pervasive problem, however, is the requirement of sets of genes that are clear orthologs across most of the species considered. This results, inevitably, in fewer genes being suitable for analysis as the number and diversity of the species considered increases. For instance, a 191-species tree (including 23 eukaryotes) was reconstructed by concatenating 31 genes (Ciccarelli et al., 2006), which raised criticism as it was not considered a fair representation of the whole genomic signal (Dagan and Martin, 2006). Limited gene sampling is especially worrying when the selected set is enriched in few functional classes, because specific footprints of selection may bias the reconstruction. In the context of such limitations, current efforts focus on increasing either gene or taxon sampling, although both approaches clearly improve the analysis by alleviating sources of phylogenetic errors (Rokas and Carroll, 2005). To overcome such limitations and to enable the efficient use of a growing number of genomes, we have devised an iterative procedure that optimizes both taxon and gene sampling at each tree partition (see Figure 1, and Material and Methods section). In brief, our procedure (NPR- for *Nested Phylogenetic Reconstruction*) starts by

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

reconstructing a standard concatenation-based tree, which is subsequently divided into two partitions by splitting a given branch. The phylogeny of the species in each resulting partition is then re-evaluated separately and further partitioned, in an iterative process ending at the desired level of resolution. As each successive step involves fewer species of higher evolutionary relatedness, the number of genes suitable for analysis is bound to increase, as is the expected quality of the resulting topologies.

We initially tested this approach on a set of 216 completely-sequenced eukaryotic genomes and show that the resulting topology is highly resolved and more accurate than a standard concatenation-based approach used over the same sets of species. In addition, our partitioned approach paves the way for subsequent updates of specific partitions of the tree, and we show this by adding, in two consecutive steps 29 and 178 newly-sequenced species, respectively.

Results and Discussion

Overview of the NPR approach

The first iteration in NPR consists of the standard procedure in a concatenation-based approach: genes present as single orthologs in most of the species considered are selected, aligned, and concatenated into a single data matrix which will constitute the input for the chosen method for phylogenetic reconstruction. Here, we opted for a blast-based approach to select sets of single-copy orthologs, which were aligned and used to reconstruct a phylogeny using a Maximum Likelihood (ML) approach, as implemented in RAxML (Stamatakis et al., 2005), using a partitioned dataset in which each concatenated gene followed the best-fitting model out of four possible ones, and using

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

four rates categories (see Materials and Methods for further details). The result of such first iteration, equivalent to a standard concatenation-based approach, was used as a reference to evaluate potential improvements of the NPR approach (see below). Subsequently, a tree partition is chosen to split the species set into two complementary subsets, which will be analyzed individually using a concatenation approach only differing to the one explained above in that it is applied only to the subset of species in that partition. The number of species considered in each split is bound to decrease as we move to more terminal branches. Similarly, if the early split is chosen close to the real root of the tree, the considered species will consist of smaller groups of increased phylogenetic relatedness and, therefore, the number of suitable genes is expected to increase. Of note, this latter expectation will not be fulfilled if a clearly wrong early split is chosen, which in turn argues for using this parameter to monitor the appropriateness of the early split selection. Finally, the resulting sub-trees from the different iterations are assembled into a single tree, whose branch lengths are re-computed using the concatenated alignment from the first iteration (containing sequences that are present in all species), so that the final branch lengths are directly comparable across partitions. The specific methods and parameters used for the different steps of the pipeline, namely i) construction and selection of orthologous groups, ii) multiple sequence alignments, and iii) phylogenetic reconstruction, can be altered within the NPR framework. Indeed, the NPR approach enables the combination of different methods and parameters at each iteration (see Material and Methods and supplementary figure S1 for additional details). Thus, we will not put an emphasis on our particular choices in the implementation of NPR, but rather on the effect of using NPR versus a single-step concatenation approach.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

A nested phylogenetic approach renders a highly resolved eukaryotic tree of life

We applied our newly-developed strategy to 216 fully-sequenced eukaryotic species, which involved 76 iterations. The particular sets of methods and parameters used in the two implementations of NPR used here are described in the material and methods section. In a first step, in order to explore the possible effect of choosing alternative splits to initialize the process, we performed 21 runs of the NPR approach using a fast implementation of the pipeline (see Materials and Methods), each one starting from an alternative earliest split. Our results (supplementary figure S2), show that most runs converged into a highly similar topology except for two cases (*Homo sapiens*, and Afrotheria splits), which resulted in highly divergent final topologies. Of note, these two splits were later found to belong to two highly unstable clades in a full NPR run (see below). Thus, the comparison of several runs performed with a fast implementation of NPR starting from alternative early splits served to inform the choice of the initial split.

We selected the branch separating all viridiplantae from the rest as a first split. This represents a clear monophyletic clade that is likely close to the root of the eukaryotic tree (Keeling et al., 2005), and was among the early splits shown to produce a robust topology in the analysis described above. We thus ran the NPR approach using a more standard and computationally-demanding phylogenetic reconstruction pipeline (see Material and Methods). Consistent with the above mentioned expectation of increased gene sampling through NPR iterations, the number of concatenated genes ranged from 131 at the deepest node to 9,525 at the node containing 8 *Drosophila* species comprising the melanogaster/obscura groups. Positive effects of the increased gene

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

5

sampling at each iteration are clear, both in terms of accuracy, as judged from the overall agreement with taxonomic classifications and established relationships (Figure 2a-c), and in terms of balanced functional representation (Figure 3a). The final topology (Figure 4, see interactive version at http://tol.cgenomics.org/euk_01) is highly resolved, with all but 6 branches in the tree receiving the highest statistical support as inferred from approximate likelihood ratio tests (aLRT). We also assessed the level of topological variation of inferred partitions by reconstructing a phylogenetic tree for each of the 226,472 alignments used in all iterations. The level of congruence with individual gene phylogenies (*i.e.* gene tree support) was computed for each TOL node by comparing with the topologies of the trees derived from the individual alignments among those comprising the super-matrix used to compute that specific node (Figure 4). Confirming earlier observations (Marcet-Houben and Gabaldón, 2009), low congruence values are present also in highly statistically supported branches, indicating the potential existence of phylogenetic noise or alternative signals such as incomplete lineage sorting and lateral gene transfer (Ané et al., 2007; Castresana, 2007; Degnan and Rosenberg, 2006; Huerta-Cepas et al., 2007).

We next investigated the congruence of taxonomic divisions, as established in NCBI, with our final topology. We note that NCBI taxonomic resource does not contain the most up-to-date and accepted taxonomic classifications, but is nevertheless a manually-curated taxonomic database which is both comprehensive and amenable for large automatic comparisons. Agreement with NCBI taxonomic divisions is remarkably high, considering our completely automated and uninformed approach. Indeed the final topology recovers the monophyly of 259 out of the 278 NCBI-based taxonomic groupings with two or more species in the tree, while the standard super-matrix

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

procedure recovered 232. Some of the observed inconsistencies are due to a few clear misplacements in our tree, including the positions of *Vitis vinifera* –expected to cluster basal to other rosids and not with *Populus*-; *Physcomitrella patents* –expected to be the earliest branching lineage in Streptophyta and not grouped with *Selaginella* -; or *Entamoeba hystolytica* and *E. dispar* –expected to be grouped with *Dictiostelium* and not with other fast-evolving parasites-, which have been problematic in earlier studies (Burleigh et al., 2011; Parfrey et al., 2010). However, most inconsistencies correspond to currently debated-clades (table S2). For instance, 10 inconsistencies are related to the use of morphology-based criteria in fungi that have recently been challenged by molecular analyses (McLaughlin et al., 2009). Additionally, our reconstruction was consistent with the fungal phylogenies published by the AFTOL project (Hibbett et al., 2007; McLaughlin et al., 2009). Similarly, 3 inconsistencies are due to the recovery of *Toxoplasma gondii* next to *Plasmodium* and *Theileria*, which is in agreement with recent molecular analyses (Kuo et al., 2008), but clashes with the classical grouping of Eimeria, Eucoccidiorida and Coccidians. Furthermore, our tree recovers nematodes as the closest relative of arthropods among the species in our analysis, thus providing support for the ecdysozoa hypothesis, grouping animals that shed their exoskeleton (Aguinaldo et al., 1997). This is in line with most recent analyses (Dunn et al., 2008), and in contrast to the alternative grouping of arthropods and chordates to the exclusion of nematodes (i.e coelomata hypothesis), which received some support in the past (Ciccarelli et al., 2006; Telford, 2004). Our tree also recovers most established clades of microbial eukaryotes such as alveolates or stramenopiles, but provides no support for the currently debated chromalveolate hypothesis joining these two groups (Keeling et al., 2005; Parfrey et al., 2010). Our tree recovers Microsporidia within fungi, a relationship that is generally elusive in phylogenetic analysis (Capella-Gutierrez et al.,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

2012). With respect to the unresolved nodes within placental mammals (Song et al., 2012), our tree supports the Afrotheria (elephant, tenrec, and hyrax in our tree) as the first branching group in the mammalian clade, followed by Xenarthra (armadillo, sloth), Laurasatheria and Euarchontoglires (glires and primates). Within the latter, our tree groups the tree shrew *Tupaia belangeri* with glires (rodents and lagomorphs) rather than with primates, as has been observed in other studies (Hallstrom and Janke, 2010). Remarkably, concatenation seems robust to the presence of low-coverage vertebrate genomes, which have been shown to introduce artefacts in gene phylogenies (Milinkovitch et al., 2010). Within arthropods, our tree supports the established phylogeny of sequenced species, including the genus *Drosophila* (Clark et al., 2007). Interestingly, the proposed (Pollard et al., 2006) incomplete lineage sorting at the speciations of *D. melanogaster*, *D. erecta* and *D. yakuba* is consistent withthe observed low level of gene tree support (0.54). Thus, our parallel computation of gene trees provides the means for pointing out possible cases of such events, and reinforces earlier proposals for including gene tree supports in phylogenomic analyses (Ané et al., 2007; Marcet-Houben and Gabaldón, 2009). Of note incomplete lineage sorting is not the only possible source of discordance between gene trees and species trees. Horizontal gene transfer, recombination, hybridization or introgression, are other biological processes that may render discordant gene trees (Degnan and Rosenberg, 2006). Thus further analyses would be necessary to disentangle the potential origins for low gene tree support at the different nodes.

Branch length analysis in the composite tree

Given the use of different gene sets, branch length estimates in the composite tree are not directly comparable. We thus re-scaled the tree by re-computing branch lengths in

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

the final topology using the 131 orthologous groups from the first iteration. Notably, we found a high correlation between the two measures ($R=0.87$ $p=6\cdot10^{-136}$), but a clear deviation towards higher values in the composite tree (supplementary figure S3). Indeed, the total length in the composite tree increased by 33%, suggesting that more widespread genes tend to evolve slower. Several explanations for this trend are possible including the difficulty of detecting distant orthologs in fast-evolving genes, or potential effects in the selection of widespread families of lineage-specific duplications followed by acceleration of one of the paralogous lineages. Our analysis of branch lengths also supports and expands previous findings (Ciccarelli et al., 2006) of a general tendency of eukaryotes with smaller genomes to evolve faster (Figure 3b). However, our broader sampling of species includes notable exceptions to this trend, including *Trichomonas vaginalis*, which constitutes the first sequenced eukaryote with a large genome (59,679 genes) that evolves significantly fast (6th in the rank), perhaps as a result of a recent, retrotransposon-mediated, expansion of its genome (Carlton et al., 2007). Microsporidian parasites and *Giardia lamblia* were found as the fastest evolving eukaryotes. We also found a strong correlation between the distance between two species and the fraction of genes they share (Figure 3c). Notably, for similar levels of shared gene content, smaller genomes tend to be at larger distances from their partners, indicating that genome reduction is associated with increased evolutionary rates. Finally, similar to what was described for eukaryotic and prokaryotic taxonomic classification (Ciccarelli et al., 2006), we found a higher level of taxonomic resolution in metazoa as compared to fungi. That is, for a given level of taxonomic classification a metazoan clade will include less divergent species as compared to a fungal clade, likely owing to our bias in assessing a greater diversity in the former (Figure S4).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

234 ***Incremental additions in Nested Phylogenetic Reconstruction***

235 A pressing challenge for the use of genomes to resolve the TOL, is the need to cope
236 with the massive production of new sequences, especially after recent technological
237 developments. In nested phylogenetic reconstruction the addition of new species does
238 not require re-computation of the whole tree, but rather of only the affected partitions.
239 Although it is difficult to have a prior knowledge of how widespread will be the effect
240 on the whole tree of adding a few extra species, our previously-described analysis of 20
241 alternative initial partitions suggest that many partitions are expected to remain stable.
242 Optimal strategies for expanding the tree and minimizing the number of computations
243 include adding sets of related species or bypassing nodes that are highly stable. Our data
244 show that nodes likely to be unaltered are highly predictable from support data obtained
245 at earlier iterations. For this, gene tree supports were more informative than statistical
246 supports, showing that branches with a gene tree congruence higher than 60% were
247 never altered in successive iterations. To put this on test, we expanded our tree by
248 adding 29 newly sequenced fungal species (supplementary table S3), which resulted in
249 the second version of our growing TOL (see http://tol.cgenomics.org/euk_02). Besides
250 the partitions including the fungal clade, only one additional basal partition (the one
251 including the long-branching unicellular parasites) changed and needed to be
252 recomputed. This result shows the suitability of NPR for incremental additions of new
253 taxa. An additional incremental step including 178 additional diverse genomes has been
254 started as we write this manuscript. This will result in a NPR-based tree of eukaryotes
255 including 418 species with complete genomes (current version available at
256 http://tol.cgenomics.org/euk_03).

257

258

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

Concluding remarks

We have proposed a novel strategy that enables the refinement of standard concatenation-based approaches by iteratively re-sampling marker genes and re-computing phylogenetic relationships. This strategy is specially suited for the automated reconstruction of species relationships when large datasets of fully-sequenced genomes are available. Admittedly, the current set of fully-sequenced species can be considered a relatively sparse and biased sampling of the global eukaryotic diversity, especially when compared to focused studies that target a few marker genes in broader sets of species (James et al., 2006; Parfrey et al., 2010; Regier et al., 2010). Nevertheless, this difference is set to diminish given the increasing rates at which new genomes are sequenced, and in the context of a growing amount of large-scale sequencing projects targeting the diversity of specific eukaryotic groups (Genome 10K Community of Scientists, 2009; Martin et al., 2011). This, coupled with the benefits of using complete genomes for phylogenetic inference (Delsuc et al., 2005; Rokas and Carroll, 2005; Rokas et al., 2003), underscores the necessity for endeavours such as the one presented here. Ideally, an initial, automatically-generated evolutionary framework provided by a method such as ours, could be later refined at specific nodes by more detailed analyses, or by incorporating additional data from ESTs or unfinished genomes.

Such automated approach should not be viewed as undermining the importance of careful and detailed analyses carried out with extensive expert curation, since the latter will always be better positioned to control for specific biases such as long-branch attraction, heterotachy, or compositional biases at particular nodes.

In particular, we have observed that it is at the earlier splits where the advantage of our strategy with other approaches is less clear. In these partitions phylogenies are based on

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

fewer genes which, as shown in this work, evolve faster. These splits would be better resolved by targeted strategies that specifically tackle known problems affecting these nodes, such as long branch attractions and horizontal gene transfer (Gribaldo and Philippe, 2002; Katz, 2002; Parfrey et al., 2010).

Finally, there is current debate on whether a tree can readily represent the true evolutionary relationships among genomes (Koonin, 2009). Indeed, processes such as horizontal gene transfer or hybridization, for instance, may be best represented by non-binary relationships such as networks. The existence of such processes, however, is still largely compatible with underlying tree-like structures corresponding to the dominant signals (Bininda-Emonds, 2005; Burleigh et al., 2011; Puigbo et al., 2009). Our focus on eukaryotes and the search of widespread genes was intended to minimize the impact of LGT in our reconstruction. Admittedly, LGT may still be an issue for the early diverging clades and additional filters to avoid the use of gene trees with largely incongruent histories may be recommended. Our combined reconstruction of species trees from concatenated alignments as well as thousands of individual gene trees provides the means not only to assess the dominant evolutionary relationship underlying the genomic data but also to identify those nodes where alternative signals are present. In addition, these gene tree collections could be directly used to derive super-trees (Bininda-Emonds, 2004), a strategy that may be preferred in some contexts. Indeed, the nested nature of our approach enables the use of different phylogenetic reconstruction strategies at different nodes in the tree.

Altogether we have presented a new phylogenetic reconstruction strategy and have explored its main limitations and advantages. While conflicting nodes, mostly those

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

close to the root of the eukaryotic tree, remain challenging and would be better dealt with by alternative approaches, NPR has shown to be an efficient approach to accurately resolve most partitions in the eukaryotic tree in a fully-automated manner. In addition NPR provides an entry point for the efficient incremental additions of new species to existing phylogenies. Given the suitability of NPR to hybrid designs that, for instance, could solve different tree splits using different methods and datasets, we believe that a sensible approach would be to incorporate NPR in the resolution of partially-constrained trees in which problematic nodes have been solved by specific approaches.

Material and Methods

Sequence data

Sequences were downloaded from various public repositories (see supplementary table S1). In all cases, whole-genome protein sequence data (i.e. proteomes) were retrieved, parsed, and stored in a local relational database.

Genome comparisons and construction of orthologous groups.

Best Reciprocal Blast Hits (Huynen and Bork, 1998) were computed for all pairs of proteomes using a Blast (Altschul et al., 1990) approach (evalue <= 0.001). Next, for every set of species defined by the internal nodes of the TOL, a collection of Orthologous Groups (OGs) was defined by finding clusters of genes that were all best reciprocal blast hits across the species considered.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

Selection of orthologous groups

At each iteration, a set of OG was selected for phylogenetic analysis. For this we ranked sets of OG by maximizing three different criteria considered important for a balanced representation of the species considered: i) average number of species represented in each OG (*A*), ii) number of OG containing the least represented species, and iii) total number of OG included in the set.

Multiple sequence alignments

Sequences in each OG were aligned using Muscle v3.6 (Edgar, 2004) with default parameters. To remove poorly aligned regions, Multiple Sequence Alignments (MSAs) were trimmed with trimAl v1.3 (Capella-Gutierrez et al., 2009) using a gap threshold of 0.1.

Nested Phylogenetic Reconstruction (algorithm)

The Nested Phylogenetic Reconstruction method addresses the analysis of every node within a precomputed phylogeny as an independent phylogenetic problem. Thus, starting from the complete set of species, , multiple hierarchical iterations are executed to optimize the topology of internal partitions. The algorithm consists of the following steps (see supplementary figure S1for the algorithm flowchart):

1. A starting unrooted tree is reconstructed including all species of interest and using a standard super-matrix approach with the preferred methodology and parameters.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

2. The starting tree is rooted using predefined and well supported monophyletic outgroup (the plants clade, in our example) and split into the resulting daughther partitions (referred here as target partitions).

3. Next, species in each of the target partitions are extracted. A set of out-group species (4 species in our case) are selected from the sister partition.

4. A new round of phylogenetic reconstruction is then executed for each of the merged sub-groups of species, including a new phase of orthology detection and specific adjustment to the phylogenetic workflow. Note that, although different workflows and approaches could be automatically applied to different nodes depending on its size or intrinsic characteristics, in our example the same pipeline was maintained for all the iterations.

5. The two resulting sub-trees are subsequently rooted using their corresponding external species, pruned, and assembled to the original main tree. While the branch length and support value for the target node are kept as observed in its parent iteration, branch information of the sibling nodes refer to the subtree obtained in step 4.

6. Finally, if any of the two major partitions observed in the resulting sub-trees contains more than a given number of species (6 in out example), they are used to feed a new iteration starting from step number 3.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

This algorithm has been implemented on top of our own computational resources at the lab. Scripts used and a beta version of a general implementation of the pipeline can be found at https://github.com/jhcepas/npr .

Gene tree reconstruction and evolutionary model selection

For each OG alignment, a model selection step was performed to choose the best fitting among 6 competing models (JTT, WAG, LG, Blosum62, VT, RtREV). For this, the likelihood of each model was computed on a topology obtained by a neighbor joining (NJ) approach, including branch length optimization as implemented in PhyML 3.0 (Guindon et al., 2010). Best fitting models were selected according to the AIC criterion (Akaike, 1973). This model-selection procedure has been used previously and has been shown to be highly accurate(Huerta-Cepas et al., 2011). Next, a Maximum Likelihood (ML) tree was reconstructed for every MSA using the best fitting model as implemented in the RAxML program (Stamatakis et al., 2005) (vesion:7.2.6, using GAMMA distribution and the rapid hill climbing algorithm). A total of 226,472 gene trees were computed.

Phylogenetic Reconstruction of TOL partitions

NPR can be used under a diversity of phylogenetic methods and specific implementations. What follows is a description of our particular choices for the discussed example. For the combined phylogenetic reconstruction, relevant trimmed MSAs were concatenated into a single alignment. RAxML (vesion:7.2.6, using GAMMA distribution with four categories and the rapid hill climbing algorithm was used to compute a ML tree using each concatenated alignment. The best fitting models of the different alignments were used in the reconstruction by defining the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

409 corresponding partitions in the concatenated alignment. To avoid over-parametrization,
410 this was done by grouping all genes with the same preferred model into a single
411 partition. Branch lengths were computed using joint estimation. The monophyly of the
412 four out-group species in each tree was constrained (see below). A fast implementation
413 of the pipeline using FastTree instead of RAxML is described below.
414
415
416 *Tree split and Subtree rooting*
417 After the reconstruction of each TOL node, two new partitions were defined according
418 to the first split of the tree topology obtained. If any of the resulting partitions was
419 found to contain more than 6 species, a new refinement step was carried out. In order to
420 allow the correct assembly of deeper nodes to their parents, 4 out-group species were
421 added to the partition, thus providing an anchoring point for subsequent rooting. Out-
422 groups were automatically selected from the sister partition considering their average
423 branch distance to the target partition. Moderately distant species were prioritized over
424 closer and farther species. This is, out-group species were selected among the species
425 whose distances were closest to the mean, rather than in the extremes of the distribution
426 of distance values. If the sister partition contained fewer than 4 species, this number was
427 completed by adding the closest species in the parent partition. The ETE toolkit
428 (Huerta-Cepas et al., 2010) was used to implement all tree operations.
429
430 *A fast implementation of NPR: FastTree phylogenetic workflow to measure the effect of*
431 *different basal rootings*
432 In order to measure the effect of different splitting strategies at the first NPR iteration,
433 we performed a series of NPR executions differing only in the earliest split selection.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

Midpoint selection of the early split and twenty manually-selected nodes generating early splits of different sizes were tested, namely: Aves, Laurasiatheria, *Capsaspora*, Fish, *Xenopus*, Human, Afrotheria, Primates, Midpoint, Dothidea, Saccharomycotina, Plants, Nematods, Drosophila, *Drosophila melanogaster*, *Entamoeba*, Alveolata, *Schizosaccharomyces*, Basidiomycota, Euglenozoa and Microsporidia. The following phylogenetic pipeline was used in all cases: Clusters of orthologous groups were selected using the same procedure as in the main pipeline. Orthologous sequences were aligned using Mafft (Katoh and Toh, 2008) with default parameters. Columns containing more than 90% gaps were removed using trimAl (Capella-Gutierrez et al., 2009). Trimmed alignments were concatenated at every iteration and used to reconstruct a tree using FasTree v2 (Price et al., 2010) under the JTT model.

Branch length optimization of the final tree

Using the topology of the final TOL, we computed a joint branch-optimization test using the basal concatenated alignment, including 131 OGs. For this, we used RAxML (version 7.2.8, with the -f e option enabled. i.e. optimizing model and branch lengths for given input tree under gamma distribution). Best fitting models for the different regions in the concatenated alignment were also supplied to the program for better optimization of branch lengths.

TOL analyses

Gene tree support

A value of gene tree support was calculated for every branch in the final tree based on the level of congruence with gene tree phylogenies reconstructed for each OG in the super-matrix used to reconstruct that partition. Thus, gene tree support for internal

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

partitions was calculated as the fraction of individual gene trees in the parent node that supported that clade. A high-resolution bubble-tree-map image showing the distribution of these values across the different tree branches can be found at http://tol.cgenomics.org/euk_01_gallery#supports

aLRT support

aLRT non-parametric branch support based on a Shimodaira-Hasegawa-like procedure were computed for every node as implemented in RaxML 2.7.8 (Stamatakis et al., 2005). Such values were calculated for every partition in the final tree using the alignment and topology of the corresponding sub-tree. Branches with aLRT values lower than 1.0 are indicated in Figure 4.

Branch stability

We define branch stability as the fraction of nested tree building iterations in which the partition defined by this branch is recovered. Thus, for each internal branch in the final tree, stability was calculated by counting how many times the same partition was found in previous iterations. A bubble-tree map image representing the distribution of these values across the different tree branches can be found at http://tol.cgenomics.org/euk_01_gallery#stability

Coverage over functional classes

Functional annotation for each gene in the human and yeast proteomes was derived from the eggnog database (Powell et al., 2011). The distribution of functional annotations for these protein families was analysed for every iteration including yeast or human, and compared to the genome-wide distribution of functions. The graphs in

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

Figure 2a and presented in the interactive TOL by clicking on the tree nodes (http://tol.cgenomics.org/euk_01) represent, for each functional category, the difference in the percentage of protein families that belong to that category in the selection of OGs in that node and in the whole reference genome. Functional categories included are: A: RNA processing and modification; B: Chromatin structure and dynamics; C: Energy production and conversion; D: Cell cycle control, cell division, chromosome partitioning; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; J: Translation, ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein turnover, chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Function unknown; T: Signal transduction mechanisms; U: Intracellular tracking, secretion, and vesicular transport;V: Defense mechanisms; W: Extracellular structures; Y: Nuclear structure; Z: Cytoskeleton.

Recovery of NCBI taxonomy groups

The full lineage tracks of all species included in our TOL were downloaded from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/), finding a total of 279 taxonomic groups with at least 2 representatives among the species considered (see supplementary table S2). The monophyly of such groups in the final TOL was tested using scripts based on the ETE toolkit (Huerta-Cepas et al., 2010).

Robinson-Foulds distances to reference trees

Two topological reference trees were chosen (see supplementary figure S5). The first one (S5A) depicted the evolution of fungal species as shown by Marcet-Houben et al. (Marcet-Houben and Gabaldón, 2009), in which branches identified as having a low (<50%) phylome support were collapsed. The second tree (S5B) represented the evolution of chordates as shown in ENSEMBL (Vilella et al., 2009), poorly supported branches in the literature were collapsed into multifurcations. The TOL was then traversed from the root to the outer leaves. Starting with the initial tree, at each step the traversed nodes were substituted by the newly reconstructed nodes. The resulting trees were then pruned so that they only contained leaves that also appeared in the reference tree. The Robinson-Foulds distance, as implemented in Ktreedist (Soria-Carrasco et al., 2007), between each derived TOL and the reference tree was calculated and then corrected by the number of multifurcated nodes present in the reference tree. Additionally, for each derived TOL the average number of OGs used to infer the nodes at a given iteration was computed.

Acknowledgements

We acknowledge funding from the Spanish Ministry of Economy and Competitiveness to TG (BIO2012-37161) and to JHC (Subprograma Juan de la Cierva: JCI2010-07614), and the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant agreement n. 310325. We are grateful to Martijn A. Huynen, Teun Boekhout, Roderic Guigó, and Fyodor Kondrashov for critically reading our manuscript.

Figure Legends

Figure 1.

Schematic representation of the nested phylogenetic reconstruction approach. First, a starting unrooted tree is reconstructed including all species (iteration 0, red node in panel A) and using a Gene Concatenation Methodology (GCM, panel C). GCM includes: C1) searching for groups of one-to-one orthologs (Ortholog Groups, OGs), C2) reconstruction of multiple sequence alignments of each OG, C3) phylogenetic reconstruction for each single OG, C4) concatenation of OG alignments, C5) species tree reconstruction based on the concatenated alignment. Secondly, the first resulting tree is split into two well supported clades, each of them defining a subset of species. GCM is then applied to each of the new sets of organisms, including four extra species as rooting anchors. As a result, two new trees are obtained (iteration 1, blue nodes in panel A). Subsequently, each of the new sub-trees is rooted using their anchor species (C6) and split into its two major clades (C7). The four resulting partitions (iteration 2, green nodes in panel A) are used to continue the same procedure until reaching a given limit for the size (number of species) in the recomputed partitions (panel B). An animation showing how the tree is re-shaped at each iteration can be seen at http://tol.cgenomics.org/TOL_animation.gif.

Figure 2

TOL analyses I: A-B) Grey lines represent topological distance between reference trees and the TOL (A-Chordates, B-Fungi, see Figure S5). Black line represents the number of protein families used at each iteration. C) Number of NCBI taxonomic groups not recovered at each iteration.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

559 **Figure 3**

560 TOL analyses II: A) Bars represent differences in percentage of protein families in
561 different functional categories for proteins used in the TOL and the Human genome at
562 three tree iterations: first node, base of chordates, and last iteration within primates. B)
563 Genome size versus distance to the root of the TOL, which was arbitrarily placed at the
564 base of metazoan+fungi+dictiostellium clade (Keeling 2005). Blue circles represent
565 metazoa, red diamonds fungi and yellow squares other eukaryotes. The yellow
566 shadowed point represents *Trichomonas vaginalis* while the grey shadowed points
567 represent Microsporidia and *Giardia lamblia*. F) Pearson Correlation between the gene
568 content score (shared genes over the minimum size of the proteomes compared) and the
569 branch length distance between the two species. Gene content scores are shown in
570 yellow for pairs including at least one small genome (<5,000 genes), red for pairs in
571 which the smallest genome is large (>16,000 genes), and blue for intermediate genome
572 sizes.

573

574 **Figure 4**

575 Representation of the final 216-species eukaryotic Tree of Life obtained by applying a
576 nested phylogenetic reconstruction: nodes represented as coloured circles indicate
577 partitions resolved using a maximized set of orthologous groups. Green nodes received
578 maximal statistical support (SH-like approximate Likelihood Ratio Test -aLRT- support
579 = 1.0), red nodes indicate aLRT statistical supports lower than 1.0. Resolution limit was
580 set to six species and partitions containing fewer than six species were not optimized.
581 Non-optimized nodes are represented as small squares and follow the same coloring
582 system to represent aLRT support. The size of the blue bubbles over internal nodes
583 indicates the fraction of gene trees supporting the monophyly of such partition (gene

584 tree support, see supplementary methods). A high resolution version of this and other
585 TOL images can be found at http://tol.cgenomics.org/gallery. All tree representations
586 were produced with ETE (Huerta-Cepas et. al. 2010).
587
588
589 **Literature cited**

590 Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff,
591 and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other
592 moulting animals. Nature 387:489-93.
593 Akaike, H. Year. Information theory and extension of the maximum likelihood principle
594 *in* Proceedings of the 2nd international symposium on information theory,
595 Budapest, Hungary:267-281.
596 Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local
597 alignment search tool. J Mol Biol 215:403-10.
598 Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation
599 of concordance among gene trees. Mol Biol Evol 24:412-26.
600 Bininda-Emonds, O. R. 2004. The evolution of supertrees. Trends Ecol Evol 19:315-22.
601 Bininda-Emonds, O. R. 2005. Supertree construction in the genomic age. Methods
602 Enzymol 395:745-57.
603 Burleigh, J. G., M. S. Bansal, O. Eulenstein, S. Hartmann, A. Wehe, and T. J. Vision.
604 2011. Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896
605 Gene Trees. Syst Biol.
606 Capella-Gutierrez, S., M. Marcet-Houben, and T. Gabaldon. 2012. Phylogenomics
607 supports microsporidia as the earliest diverging clade of sequenced fungi. BMC
608 Biol 10:47.
609 Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldón. 2009. trimAl: a tool for
610 automated alignment trimming in large-scale phylogenetic analyses.
611 Bioinformatics 25:1972-3.
612 Carlton, J. M., R. P. Hirt, J. C. Silva, A. L. Delcher, M. Schatz, Q. Zhao, J. R. Wortman,
613 S. L. Bidwell, U. C. Alsmark, S. Besteiro, T. Sicheritz-Ponten, C. J. Noel, J. B.
614 Dacks, P. G. Foster, C. Simillion, Y. Van de Peer, D. Miranda-Saavedra, G. J.
615 Barton, G. D. Westrop, S. Muller, D. Dessi, P. L. Fiori, Q. Ren, I. Paulsen, H.
616 Zhang, F. D. Bastida-Corcuera, A. Simoes-Barbosa, M. T. Brown, R. D. Hayes,
617 M. Mukherjee, C. Y. Okumura, R. Schneider, A. J. Smith, S. Vanacova, M.
618 Villalvazo, B. J. Haas, M. Pertea, T. V. Feldblyum, T. R. Utterback, C. L. Shu,
619 K. Osoegawa, P. J. de Jong, I. Hrdy, L. Horvathova, Z. Zubacova, P. Dolezal, S.
620 B. Malik, J. M. Logsdon, Jr., K. Henze, A. Gupta, C. C. Wang, R. L. Dunne, J.
621 A. Upcroft, P. Upcroft, O. White, S. L. Salzberg, P. Tang, C. H. Chiu, Y. S. Lee,
622 T. M. Embley, G. H. Coombs, J. C. Mottram, J. Tachezy, C. M. Fraser-Liggett,
623 and P. J. Johnson. 2007. Draft genome sequence of the sexually transmitted
624 pathogen Trichomonas vaginalis. Science 315:207-12.
625 Castresana, J. 2007. Topological variation in single-gene phylogenetic trees. Genome
626 Biol 8:216.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

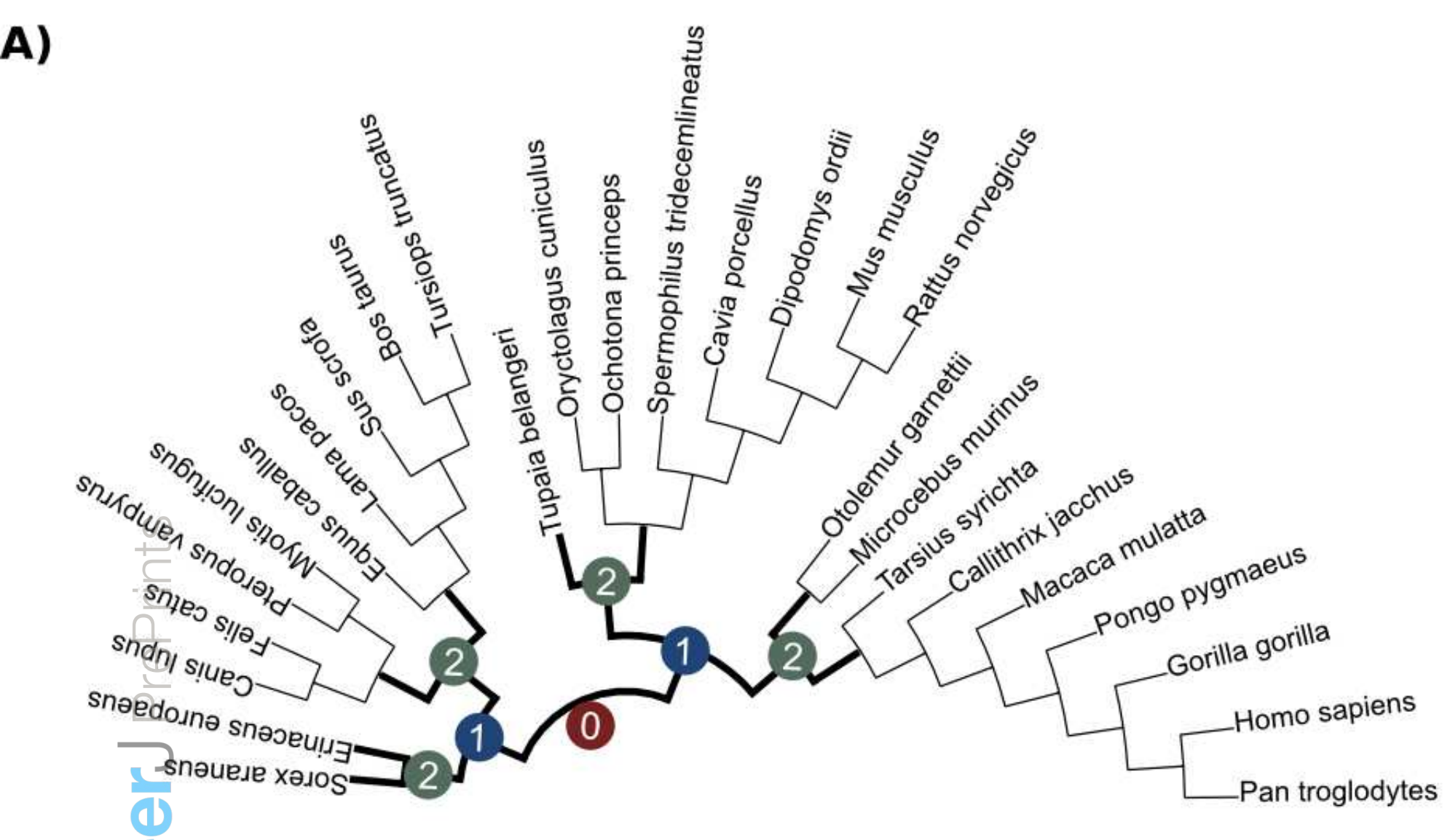
627 Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006.
628 Toward automatic reconstruction of a highly resolved tree of life. Science
629 311:1283-7.
630 Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C.
631 Kaufman, M. Kellis, W. Gelbart, V. N. Iyer, D. A. Pollard, T. B. Sackton, A. M.
632 Larracuente, N. D. Singh, J. P. Abad, D. N. Abt, B. Adryan, M. Aguade, H.
633 Akashi, W. W. Anderson, C. F. Aquadro, D. H. Ardell, R. Arguello, C. G.
634 Artieri, D. A. Barbash, D. Barker, P. Barsanti, P. Batterham, S. Batzoglou, D.
635 Begun, A. Bhutkar, E. Blanco, S. A. Bosak, R. K. Bradley, A. D. Brand, M. R.
636 Brent, A. N. Brooks, R. H. Brown, R. K. Butlin, C. Caggese, B. R. Calvi, A.
637 Bernardo de Carvalho, A. Caspi, S. Castrezana, S. E. Celniker, J. L. Chang, C.
638 Chapple, S. Chatterji, A. Chinwalla, A. Civetta, S. W. Clifton, J. M. Comeron, J.
639 C. Costello, J. A. Coyne, J. Daub, R. G. David, A. L. Delcher, K. Delehaunty, C.
640 B. Do, H. Ebling, K. Edwards, T. Eickbush, J. D. Evans, A. Filipski, S. Findeiss,
641 E. Freyhult, L. Fulton, R. Fulton, A. C. Garcia, A. Gardiner, D. A. Garfield, B.
642 E. Garvin, G. Gibson, D. Gilbert, S. Gnerre, J. Godfrey, R. Good, V. Gotea, B.
643 Gravely, A. J. Greenberg, S. Griffiths-Jones, S. Gross, R. Guigo, E. A.
644 Gustafson, W. Haerty, M. W. Hahn, D. L. Halligan, A. L. Halpern, G. M. Halter,
645 M. V. Han, A. Heger, L. Hillier, A. S. Hinrichs, I. Holmes, R. A. Hoskins, M. J.
646 Hubisz, D. Hultmark, M. A. Huntley, D. B. Jaffe, S. Jagadeeshan, et al. 2007.
647 Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-
648 18.
649 Dagan, T., and W. Martin. 2006. The tree of one percent. Genome Biol 7:118.
650 Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most
651 likely gene trees. PLoS Genet 2:e68.
652 Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction
653 of the tree of life. Nat Rev Genet 6:361-75.
654 Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver,
655 G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. Haddock, A.
656 Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q.
657 Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves
658 resolution of the animal tree of life. Nature 452:745-9.
659 Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
660 throughput. Nucleic Acids Res 32:1792-7.
661 Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-
662 genome sequence for 10,000 vertebrate species. Pages 659-74 *in* J Hered.
663 Gribaldo, S., and H. Philippe. 2002. Ancient phylogenetic relationships. Theor Popul
664 Biol 61:391-408.
665 Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel.
666 2010. New algorithms and methods to estimate maximum-likelihood
667 phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-21.
668 Hallstrom, B. M., and A. Janke. 2010. Mammalian evolution may not be strictly
669 bifurcating. Mol Biol Evol 27:2804-16.
670 Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S.
671 Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni,
672 P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W.
673 Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D.
674 Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D.
675 M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K.
676 Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

677 K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M.
678 Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D.
679 Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G.
680 Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M.
681 M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level
682 phylogenetic classification of the Fungi. Mycol Res 111:509-47.
683 Huerta-Cepas, J., S. Capella-Gutierrez, L. P. Pryszcz, I. Denisov, D. Kormes, M.
684 Marcet-Houben, and T. Gabaldon. 2011. PhylomeDB v3.0: an expanding
685 repository of genome-wide collections of trees, alignments and phylogeny-based
686 orthology and paralogy predictions. Nucleic Acids Res 39:D556-60.
687 Huerta-Cepas, J., H. Dopazo, J. Dopazo, and T. Gabaldón. 2007. The human phylome.
688 Genome Biol 8:R109.
689 Huerta-Cepas, J., J. Dopazo, and T. Gabaldón. 2010. ETE: a python Environment for
690 Tree Exploration. BMC Bioinformatics 11:24.
691 Huynen, M. A., and P. Bork. 1998. Measuring genome evolution. Proc Natl Acad Sci U
692 S A 95:5849-56.
693 James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio,
694 C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A.
695 E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B.
696 O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W.
697 Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D.
698 Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith,
699 D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D.
700 Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J.
701 Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K.
702 W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking,
703 B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R.
704 Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R.
705 Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene
706 phylogeny. Nature 443:818-22.
707 Katoh, K., and H. Toh. 2008. Recent developments in the MAFFT multiple sequence
708 alignment program. Brief Bioinform 9:286-98.
709 Katz, L. A. 2002. Lateral gene transfers and the evolution of eukaryotes: theories and
710 data. Int J Syst Evol Microbiol 52:1893-900.
711 Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J.
712 Roger, and M. W. Gray. 2005. The tree of eukaryotes. Trends Ecol Evol 20:670-
713 6.
714 Koonin, E. V. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res
715 37:1011-34.
716 Kuo, C. H., J. P. Wares, and J. C. Kissinger. 2008. The Apicomplexan whole-genome
717 phylogeny: an analysis of incongruence among gene trees. Mol Biol Evol
718 25:2689-98.
719 Marcet-Houben, M., and T. Gabaldón. 2009. The tree versus the forest: the fungal tree
720 of life and the topological diversity within the yeast phylome. PLoS One
721 4:e4357.
722 Martin, F., D. Cullen, D. Hibbett, A. Pisabarro, J. W. Spatafora, S. E. Baker, and I. V.
723 Grigoriev. 2011. Sequencing the fungal tree of life. New Phytol.
724 McLaughlin, D. J., D. S. Hibbett, F. Lutzoni, J. W. Spatafora, and R. Vilgalys. 2009.
725 The search for the fungal tree of life. Trends Microbiol 17:488-97.

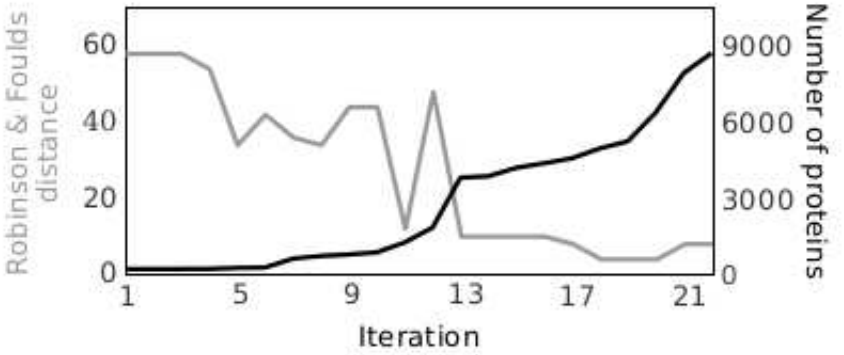
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

726 Milinkovitch, M. C., R. Helaers, E. Depiereux, A. C. Tzika, and T. Gabaldon. 2010. 2X
727 genomes - depth does matter. Genome Biol 11:R16.
728 Parfrey, L. W., J. Grant, Y. I. Tekle, E. Lasek-Nesselquist, H. G. Morrison, M. L.
729 Sogin, D. J. Patterson, and L. A. Katz. 2010. Broadly sampled multigene
730 analyses yield a well-resolved eukaryotic tree of life. Syst Biol 59:518-33.
731 Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread
732 discordance of gene trees with species tree in Drosophila: evidence for
733 incomplete lineage sorting. PLoS Genet 2:e173.
734 Powell, S., D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T.
735 Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. 2011.
736 eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different
737 taxonomic ranges. Nucleic Acids Res 40:D284-9.
738 Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2--approximately maximum-
739 likelihood trees for large alignments. PLoS One 5:e9490.
740 Puigbo, P., Y. I. Wolf, and E. V. Koonin. 2009. Search for a 'Tree of Life' in the thicket
741 of the phylogenetic forest. J Biol 8:59.
742 Regier, J. C., J. W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J. W. Martin, and
743 C. W. Cunningham. 2010. Arthropod relationships revealed by phylogenomic
744 analysis of nuclear protein-coding sequences. Nature 463:1079-83.
745 Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution
746 of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol
747 22:1337-44.
748 Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches
749 to resolving incongruence in molecular phylogenies. Nature 425:798-804.
750 Song, S., L. Liu, S. V. Edwards, and S. Wu. 2012. Resolving conflict in eutherian
751 mammal phylogeny using phylogenomics and the multispecies coalescent
752 model. Proc Natl Acad Sci U S A 109:14942-7.
753 Soria-Carrasco, V., G. Talavera, J. Igea, and J. Castresana. 2007. The K tree score:
754 quantification of differences in the relative branch length and topology of
755 phylogenetic trees. Bioinformatics 23:2954-6.
756 Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: a fast program for
757 maximum likelihood-based inference of large phylogenetic trees. Bioinformatics
758 21:456-63.
759 Telford, M. J. 2004. Animal phylogeny: back to the coelomata? Curr Biol 14:R274-6.
760 Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. 2009.
761 EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees
762 in vertebrates. Genome Res 19:327-35.
763
764

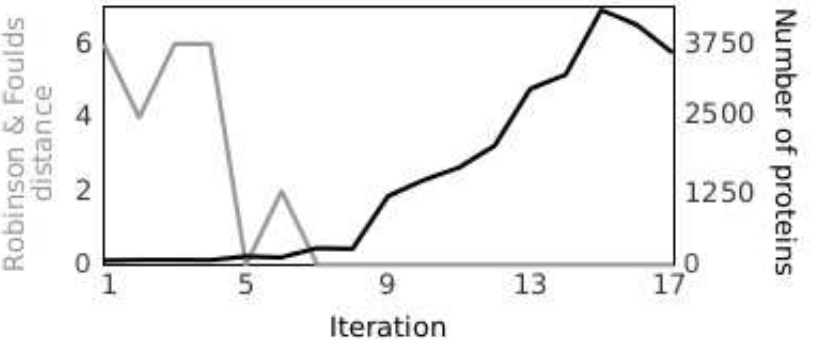


PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014

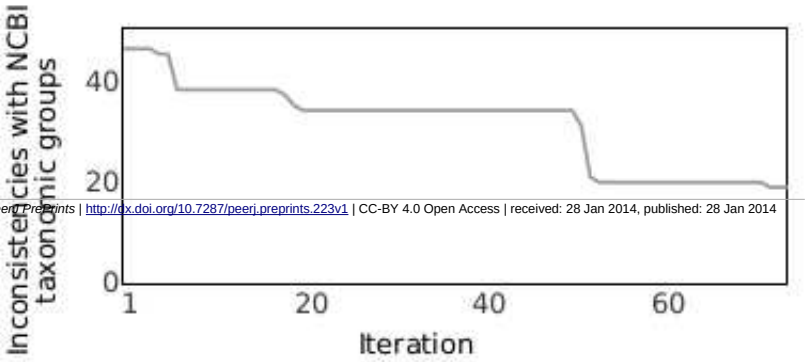
A)



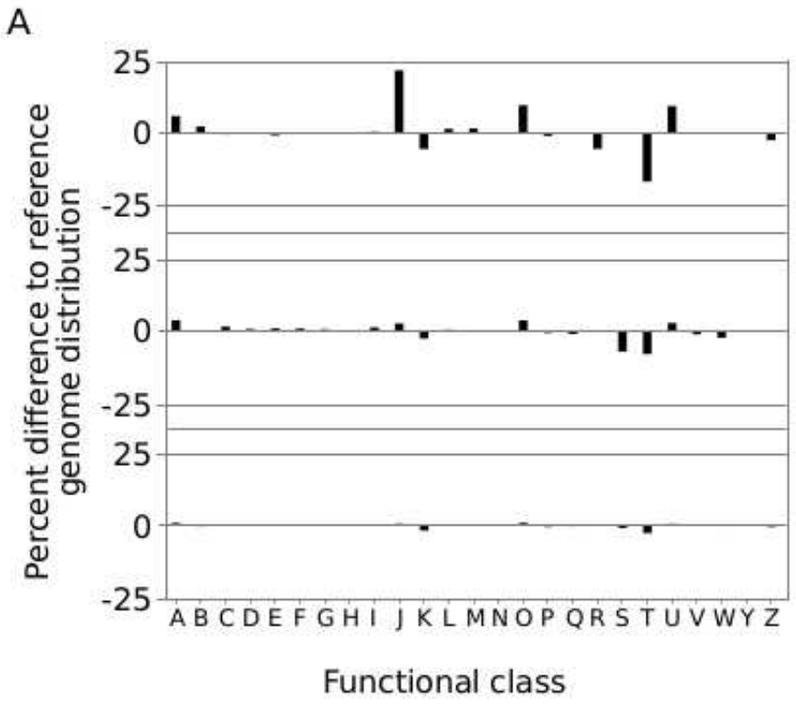
B)

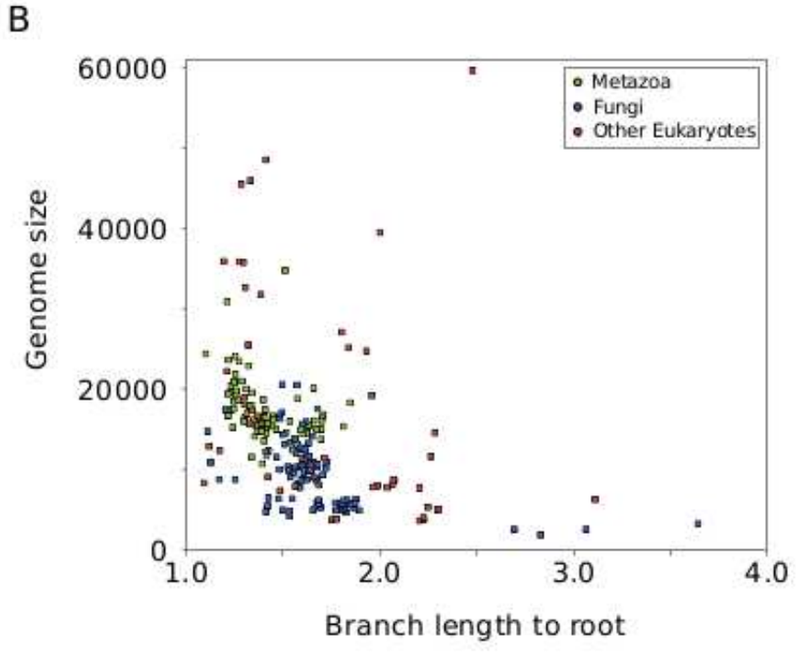


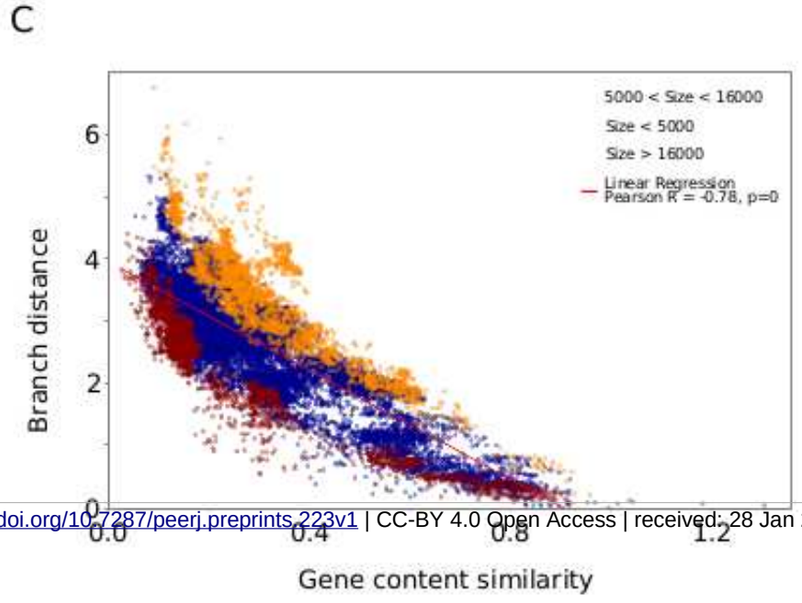
C)



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014







PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.223v1 | CC-BY 4.0 Open Access | received: 28 Jan 2014, published: 28 Jan 2014