Bioinformatics Advance Access published January 19, 2007

Original paper

Assessment of phylogenomic and orthology approaches for phylogenetic inference

B.E. Dutilh [1*], V. van Noort [1], R.T.J.M. van der Heijden [1], T. Boekhout [2], B. Snel [3] and M.A. Huynen [1]

[1] Center for Molecular and Biomolecular Informatics / Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center. PO Box 9101, 6500 HB, Nijmegen, The Netherlands.

[2] Centraalbureau voor Schimmelcultures. Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands.

[3] Bioinformatics group, Department Biology, Utrecht University. Padualaan 8, 3584 CH, Utrecht, The Netherlands.

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Phylogenomics integrates the vast amount of phylogenetic information contained in complete genome sequences, and is rapidly becoming the standard for inferring reliable species phylogenies. There are however fundamental differences between the ways in which phylogenomic approaches like gene content, superalignment, superdistance and supertree integrate the phylogenetic information from separate orthologous groups. Furthermore, they all depend on the method by which the orthologous groups are initially determined. Here, we systematically compare these four phylogenomic approaches, in parallel with three approaches for large-scale orthology determination: pairwise orthology, cluster orthology and tree-based orthology.

Results: Including various phylogenetic methods, we apply a total of 54 fully automated phylogenomic procedures to the Fungi, the eukaryotic clade with the largest number of sequenced genomes, for which we retrieved a golden standard phylogeny from the literature. Phylogenomic trees based on gene content show, relative to the other methods, a bias in the tree topology that parallels convergence in life style among the species compared, indicating convergence in gene content.

Conclusions: Complete genomes are no warrant for good, or even consistent phylogenies. However, the large amounts of data in genomes enable us to carefully select the data most suitable for phylogenomic inference. In terms of performance, the superalignment approach, combined with restrictive orthology, is the most successful in recovering a fungal phylogeny that agrees with current taxonomic views, and allows us to obtain a high resolution phylogeny. We provide solid support for what has grown to be common practice in phylogenomics during its advance in recent years.

Contact: dutilh@cmbi.ru.nl

1 INTRODUCTION

Phylogenomics, i.e., using entire genomes to infer a species tree, has become the *de facto* standard for reconstructing reliable phylogenies (Ciccarelli, et al., 2006; Daubin, et al., 2002). Whereas phylogenetic trees, i.e., based on single gene families, may show conflict (Teichmann and Mitchison, 1999) due to a variety of causes, phylogenomic trees have held the promise that they can average out these anomalies by the sheer power of genome-scale data. As it is based on the maximum genetic information, a phylogenomic tree should be the best reflection of the evolutionary history of the species, assuming this history is tree-like (Doolittle, 1999; Ge, et al., 2005). Although there are discordant processes at the level of gene repertoires, such as horizontal gene transfer (Doolittle, 1999) or differences in the rates of evolution and gene loss between paralogs in different species (Daubin, et al., 2003), these have been shown to add noise rather than a directional bias (Dutilh, et al., 2004). However, this does not mean that phylogenomics is the end of all conflict in species trees (Jeffroy, et al., 2006): there are many ways to integrate the information from the different gene families to form a single species phylogeny.

1.1 Phylogenomics

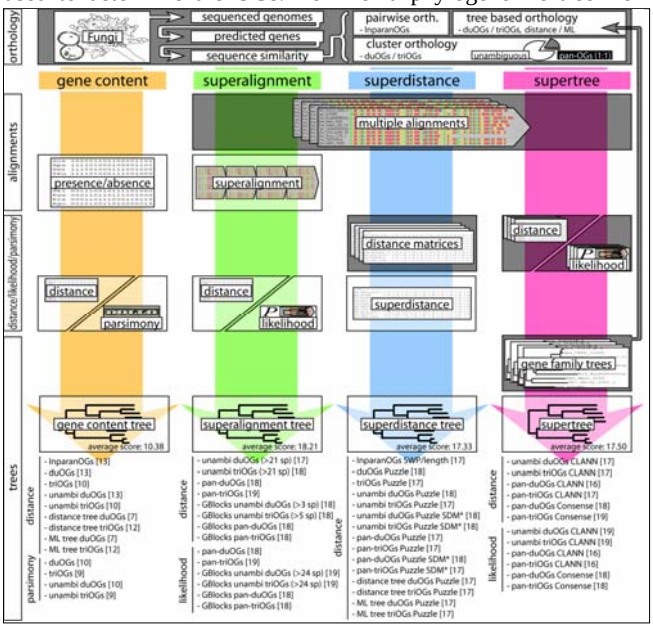
In taxonomy, the term phylogenomics indicates the construction of a phylogeny on the basis of complete genome data. We can consider this type of phylogenomics as parallel phylogenetics over all gene families, combined with a synthesis step. This step from phylogenetics to phylogenomics integrates the phylogenetic information from the different gene families to form a single species phylogeny, and can be taken at successive levels in the process. As a guide line, we classify phylogenomic methods by the level where the step from phylogenetics to phylogenomics is made (Fig. 1). Here, we compare these four qualitatively different phylogenomic approaches.

For sequence-based phylogenomic methods, the first step is to make multiple alignments for every orthologous group (OG) (Delsuc, et al., 2005). In the superalignment approach, the phylogenetic information is then combined by concatenating the multiple alignments to form a superalignment. Subsequently, conventional phylogenetic inference methods can be used to transform the alignment into a phylogeny. Superdistance trees continue the path of phylogenetics by first calculating distance matrices for all gene families. The phylogenomic distance between two species is then defined as the average distance between all the shared gene families (Kunin, et al., 2005). Finally, the supertree approach (Binindа-Emonds, 2004; Daubin, et al., 2002) takes the step from phylogenetics to phylogenomics at the very end. After phylogenetic trees have been composed for all gene families, an integration step combines the multiple gene family trees to form a single phylogenomic tree.

Of the methods based on whole-genome features (Delsuc, et al., 2005) we only consider gene content here, as gene order in the Fungi evolves too fast to retain a phylogenetic signal (Huynen, et

*To whom correspondence should be addressed.

al., 2001). Gene content takes the step from phylogenetics to phylogenomics right after the definition of the OGs (Fig. 1). Species are regarded as "bags of genes", and sequence information is only used to determine the OGs. To infer a phylogenomic tree from



gene content data, a binary character matrix indicating the presence or absence of the OGs in all species can be treated in the same way as a multiple sequence alignment.

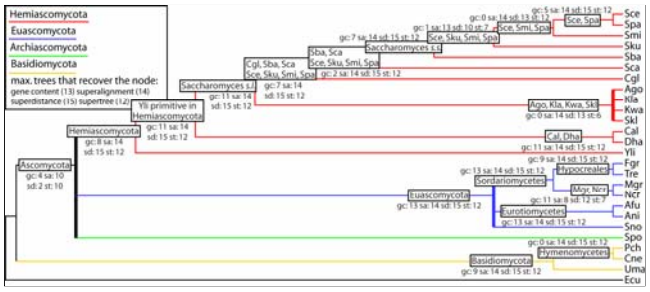
Fig. 1. Making phylogenomic trees. Before starting tree inference, OGs are defined (top row). Phylogenomics follows the steps of phylogenetics, from multiple alignment through distance, likelihood or parsimony to the reconstruction of a phylogeny. Integrating separate phylogenetics for each gene family (gray boxes) to phylogenomics (white boxes) can be done at every one of these steps. This defines the phylogenomic approach: gene content (after OG definition), superalignment (after multiple alignment), superdistance (after distance calculation) or supertree (after reconstruction of gene family trees). The phylogenomic trees we reconstructed are listed at the bottom, the number between square brackets indicates the number of target nodes that the tree recovered correctly.

1.2 Orthology

The initial step in every phylogenomic approach is to determine which genes are to be compared between species (top row in Fig. 1). We compare the performance of three types of orthology definition: pairwise orthology, cluster orthology, and tree-based orthology. The first two methods use sequence similarity scores to define orthologous groups of genes. Pairwise orthology is defined between only two species (e.g. bi-directional best hits or Inparanoid (Remm, et al., 2001)), and cluster orthology (e.g. Clusters of Orthologous Groups (Tatusov, et al., 1997)) is the natural extension of pairwise orthology to more than two species. Tree-based orthology comes closest to the original phylogenetic definition of orthology (Fitch, 1970). Rather than using only the sequence similarity scores, it analyses a phylogenetic tree of a homologous group of genes to obtain orthologous relations (van der Heijden, et al., submitted). Note that although tree-based orthology is an ideal approach to determine orthology at scalable levels of resolution, it needs to be operationalized: OGs have to be determined from the trees separately for each pair of species. The superalignment and supertree approaches, that consider a large set of species simultaneously, can not deal with pairwise orthology or operationalized tree-based orthology (see "Methods" and supp. mat.).

1.3 Fungal phylogeny

To compare the performance of phylogenomic approaches, some kind of golden standard phylogeny is imperative. We chose here to benchmark the phylogenomic methods using a phylogeny of real species. The alternative, to work with simulated evolutionary data (Hillis, et al., 1994), would require the simulation of the evolution of complete genomes for which we lack the models and parameters. *Prima facie*, an approach that uses a known phylogeny appears to exclude the possibility for any improvements. However, due to ambiguities in the literature our golden standard phylogeny is not completely resolved. We expect that properly derived com-



plete genome phylogenies will allow a higher resolution both for the species analyzed here, and for other (partly) unresolved clades in future analyses.

Fig. 2. Target phylogeny. Labeled nodes are supported by literature. Unresolved issues are indicated by multifurcating nodes (bold lines). The numbers at every node indicate the number of the trees in each of the phylogenomic approaches that recovered this node correctly. See supp. mat. for references that support this tree.

The Fungi are the eukaryotic clade with the most sequenced genomes. *Saccharomyces cerevisiae* has been a model organism for decades, and in this era of comparative genomics much work has focused on sequencing the genomes of more or less closely related species (Cliften, et al., 2003; Dujon, et al., 2004; Kellis, et al., 2003). In total, 26 completely sequenced fungal genomes were available in public databases at the start of this study (September 2005): 22 Ascomycota, 3 Basidiomycota and the Microsporidium *Encephalitozoon cuniculi* (see Fig. 2 and Table 1). We included *E. cuniculi* as an outgroup because this was the most closely related complete genome to the Fungi (Thomarat, et al., 2004; Vivares, et al., 2002), and *Rhizopus oryzae* was not available yet.

The fungal kingdom has been extensively studied by phylogeneticists. Traditional phenotypic methods (e.g. reviewed in (Guarro, et al., 1999)), molecular phylogenetic analyses based on rRNA (Fell, et al., 2000; Lopandic, et al., 2005; Lutzoni, et al., 2004; Scorzetti, et al., 2002; Tehler, et al., 2003) or small numbers of other proteins (Diezmann, et al., 2004; James, et al., 2006; Kouvelis, et al., 2004; Kurtzman, 2003), as well as some large scale studies (Jeffroy, et al., 2006; Kuramae, et al., 2006; Robbertse, et al., 2006; Rokas, et al., 2003; Thomarat, et al., 2004) have helped resolve many of the phylogenetic relationships in the fungal kingdom. Based on the available literature (Berbee, et al., 2000; Delsuc, et al., 2005; Diezmann, et al., 2004; Jeffroy, et al., 2006; Jones, et al., 2004; Kouvelis, et al., 2004; Kuramae, et al., 2006; Kurtzman, 2003; Lopandic, et al., 2005; Lutzoni, et al., 2004; Me-

dina, 2005; Prillinger, et al., 2002; Robbertse, et al., 2006; Tehler, et al., 2003; Thomarat, et al., 2004), we composed a true fungal phylogeny (Fig. 2) that we use as a benchmark.

Table 1. The organisms included in this research.

	Species name	Genes	Reference
Ago	Ashbya gossypii (Eremothecium)	4,720	(Dietrich, et al., 2004)
Afu	Aspergillus fumigatus	9,926	(Nierman, et al., 2005)
Ani	Aspergillus nidulans	9,541	(Galagan, et al., 2005)
Cal	Candida albicans	11,904	(Jones, et al., 2004)
Cgl	Candida glabrata	5,272	(Dujon, et al., 2004)
Cne	Cryptococcus neoformans	5,882	(Loftus, et al., 2005)
Dha	Debaryomyces hansenii	6,896	(Dujon, et al., 2004)
Ecu	Encephalitozoon cuniculi	1,918	(Katinka, et al., 2001)
Fgr	Fusarium graminearum	11,640	(http://www.broad.mit.edu)
Kla	Kluyveromyces lactis	5,331	(Dujon, et al., 2004)
Kwa	Kluyveromyces waltii	5,230	(Kellis, et al., 2004)
Mgr	Magnaporthe grisea	11,109	(Dean, et al., 2005)
Ncr	Neurospora crassa	10,620	(Galagan, et al., 2003)
Pch	Phanerochaete chrysosporium	11,777	(Martinez, et al., 2004)
Sba	Saccharomyces bayanus	4,966	(Kellis, et al., 2003)
Sca	Saccharomyces castellii	4,690	(Cliften, et al., 2003)
Sce	Saccharomyces cerevisiae	6,702	(Goffeau, et al., 1996)
Skl	Saccharomyces kluyveri	2,992	(Cliften, et al., 2003)
Sku	Saccharomyces kudriavzevii	3,813	(Cliften, et al., 2003)
Smi	Saccharomyces mikatae	3,100	(Kellis, et al., 2003)
Spa	Saccharomyces paradoxus	8,955	(Kellis, et al., 2003)
Spo	Schizosaccharomyces pombe	4,990	(Wood, et al., 2002)
Sno	Stagonospora nodorum	16,597	(http://www.broad.mit.edu)
Tre	Trichoderma reesei	9,997	(http://www.jgi.doe.gov)
Uma	Ustilago maydis	6,522	(Kamper, et al., 2006)
Yli	Yarrowia lipolytica	6,666	(Dujon, et al., 2004)

1.4 This study

Here, we compare the four phylogenomic and the three orthology approaches presented above (Fig. 2) in parallel, assessing their ability to infer the 19 target nodes derived from the literature. As many different methods and algorithms exist for most of these approaches, we include several implementations in order to buffer our findings from possible biases in the individual methods. Thus, we compose a total of 54 phylogenomic trees of the 26 complete fungal genomes, using completely automated methods.

2 METHODS

2.1 Orthology

Sequences were downloaded from the respective fungal sequencing projects (see Table 1). We compare the performance of three types of orthology definition: pairwise orthology, cluster orthology, and tree-based orthology. Using Inparanoid (Remm, et al., 2001), we detected 1,025,849 pairwise "InparanOGs". For cluster orthology we used a method based on COG (Tatusov, et al., 1997), yielding 8,044 triangle based "triOGs" and 10,754 pair based "duOGs". For specific purposes (supp. mat.), we composed subsets of OGs without paralogs (8,722 unambiguous duOGs and 6,488 unambiguous triOGs) and OGs that occur exactly once in every species (64 pan-duOGs and 59 pan-triOGs). To compose tree-based orthology, phylogenetic trees were analyzed with LOFT (van der Heijden, et al., submitted). LOFT does not impose a phylogeny on the data, but assigns orthology relations based on the species overlap between the branches of a phylogenetic tree. Because tree-based orthology yields levels of orthology, it needs to be operationalized between species pairs. We identified 858,622 distance tree-duOGs, 820,007 distance tree-triOGs, 856,363 likelihood tree-duOGs and 822,570 likelihood tree-triOGs. Further details about the orthology approaches can be found in the supp. mat. Orthology predictions are available at www.cmbi.ru.nl/~dutilh/phylogenomics.

2.2 Phylogenomics

Phylogenomic trees based on gene content were calculated from presence-absence profiles using either distance (Dutilh, et al., 2004; Korbel, et al., 2002) or parsimony (Farris, 1977; Felsenstein, 1989). In the distance approach, we corrected for genome size, because distantly related species with large genomes may share more genes than closer related species with small genomes (supp. mat.). For the superalignment approach, Muscle multiple alignments (Edgar, 2004) of either unambiguous cluster OGs or pan-OGs were concatenated to form a superalignment. Unambiguous OGs that are absent from certain species were coded with question marks, and form gaps in the alignment (Philippe, et al., 2004). In some superalignment trees, we analysed the effect of selecting unambiguously aligned amino acids by using GBlocks (Castresana, 2000). We used either distance or maximum likelihood approaches to reconstruct the superalignment trees. The superdistance trees were calculated from superdistance matrices, based on the average distance over all OGs that are shared between the two species. We analysed the effect of correcting for rapidly evolving OGs by using SDM* (Criscuolo, et al., 2006). Supertrees were composed of distance or maximum likelihood gene family trees. To integrate the different phylogenetic trees into a phylogenomic supertree, we used either the majority rule from Consense (Felsenstein, 1989), or CLANN (Creevey and McInerney, 2005). For further details see the supp. mat., all the trees are available at www.cmbi.ru.nl/~dutilh/phylogenomics.

2.3 Scoring the reconstructed trees

To score the reconstructed phylogenomic trees, we use the target phylogeny in Fig. 2. A phylogeny receives one point for each of the resolved partitions that is correctly retrieved, so a maximum of 19 points can be obtained. Note that, for example, the node "Yli primitive in Hemiascomycetes" refers to the (Ago, Cal, Cgl, Dha, Kla, Kwa, Sba, Sca, Sce, Skl, Sku, Smi, Spa) branch (see Fig. 2). This means that this node can contribute a point for a certain tree, even if the Hemiascomycetes are not monophyletic in that tree, for example if Y. lipolytica clusters with Sch. pombe. In that case, however, it will not receive a point for the "Hemiascomycetes" node.

3 RESULTS

We present a systematic comparison of two important factors in phylogenomic inference: the orthology approach and the level of integration of phylogenetic information to a genomic scale. We use various implementations for each of these approaches, such as the inclusive pair-based or the more restrictive triangle-based cluster OGs; and distance, maximum likelihood or parsimony for the reconstruction of the tree (Fig. 1 and supp. mat.). Thus, we automatically construct 54 phylogenies from the available genome data of 26 Fungi. To assess the performance of the phylogenomic methods, we compare the nodes in the reconstructed trees to the 19 resolved nodes of a partly unresolved golden standard phylogeny based on extensive literature research (Fig. 2 and supp. mat.). All of the canonical phylogenomic methods that we tested perform remarkably well at reconstructing the known fungal phylogeny. The phylogenomic trees in the three sequence-based approaches (superalignment, superdistance and supertree) recovered at least 16 out of the 19 target nodes. This constitutes a major distinction with the gene content trees, that performed much less well: even the

best methods recovered no more than 13 nodes. All the phylogenomic trees can be found in the supp. mat.

3.1 Collapsing recent duplications to gain data

We included two types of cluster orthology: the inclusive pair-based "duOGs", and the more restrictive triangle based "triOGs" (see "Methods"). A subset of these cluster OGs are the unambiguous OGs, that occur no more than once in every species. Even more constrained are the pan-orthologs, that are both unambiguous and universal, occurring exactly once in every species. We detected 8,722 unambiguous duOGs, 6,488 unambiguous triOGs, 64 pan-duOGs and 59 pan-triOGs in the Fungi. This result depends on collapsing the recent duplications, as identified from the phylogenies by LOFT (van der Heijden, et al., submitted), before selecting the unambiguous OGs from the cluster OGs (see supp. mat.). Without collapsing recent duplications, we retrieved no more than 4,421 unambiguous duOGs, 4,887 unambiguous triOGs, 13 pan-duOGs and 13 pan-triOGs. This difference (an average of 42%) illustrates the necessity to filter out species-specific gene expansions and systematic errors, such as the diploid genome assembly of *Can. albicans* (Jones, et al., 2004), to increase the number of genes that can be considered.

3.2 Orthology approaches

An orthology definition that considers a recent last common ancestor will have a higher resolution than one that considers a more ancient common ancestor. Thus, pairwise orthology and tree-based orthology should, in principle, obtain a higher resolution than cluster orthology, that includes in a single OG all gene duplications since the last common ancestor of all the species compared. However, pairwise orthology incorporates information from only two species, and may miss genes that cluster orthology and tree-based orthology can identify. We expected tree-based orthology, that includes sequence information from many different species, while allowing a high-resolution view where necessary, to combine the advantages from pairwise and cluster orthology. However, although the orthology definition does turn out to be an important factor in the quality of a phylogenomic tree, the highest scoring trees were based on either unambiguous cluster OGs (duOGs and triOGs) or pan-triOGs, rather than tree-based OGs.

It is striking that although there is a large overlap between the 64 pan-duOGs and 59 pan-triOGs (56 OGs are identical), the pan-triOGs give better trees in both the superalignment and the supertree approach. However, the choice for one of these orthology definitions is no warrant for a good phylogeny. Both the unambiguous cluster OGs and the pan-triOGs also produced relatively low-scoring trees in every phylogenomic approach (Fig. 1).

3.3 Superalignment trees and supertrees can recover all target nodes

Superalignment can be considered the most successful phylogenomic approach: four of the 14 superalignment trees correctly infer all 19 target nodes (see Fig. 1). The most difficult to recover as a monophyletic group are the Ascomycota (although not for the trees constructed with maximum likelihood) and the (Mgr, Ncr) node (Fig. 2). In those superalignment trees that did not group *M. grisea* with *N. crassa*, neither of these species was preferentially found at the root of the Sordariomycetes.

Selecting the unambiguously aligned positions of the superalignment using GBlocks (Castresana, 2000) made it computationally possible to include more unambiguous OGs (supp. mat.), which led the unambiguous duOGs to match the results of the unambiguous triOGs (Fig. 1). However, the decrease in the number of aligned positions that GBlocks brought about in the pan-triOGs, resulted in a sub-optimal tree (Fig. 1). It appears that it is not simply the selection of unambiguously aligned positions, but rather the increase in the amount of high quality data that leads to a better phylogeny. To further test this, we composed Consense supertrees from an increasing number of phylogenetic distance trees of the most restrictive OG set, the 59 pan-triOGs. Interestingly, no two single gene trees were identical, and none was identical to the target: on average, they recover only 11.5 nodes. Yet when we combine at least 30-40 phylogenetic trees to a supertree, we already recover the external golden standard (*Figure 3* in the supp. mat.).

Three of the 12 phylogenomic trees inferred using the supertree approach correctly recover all 19 target nodes. The Consense supertree based on phylogenetic distance trees from pan-triOGs is identical to the four highest scoring superalignment trees (Fig. 3a), but differs slightly from the equally high-scoring Clann supertrees based on phylogenetic maximum likelihood trees from both duOGs and triOGs (supp. mat.). This is possible because of the unresolved nodes in the target phylogeny. Note that superdistance and gene content trees never retrieve all 19 target nodes.

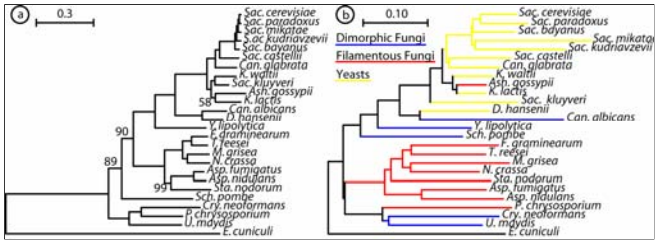


Fig. 3. Phylogenomic trees. a) One of the two highest scoring fungal topologies. This topology was recovered by four superalignment trees and one supertree. A ML tree based on a superalignment of pan-triOGs, a ML tree based on a GBlocks-filtered superalignment of unambiguous duOGs (present in >24 species, 132,409 positions; this is the tree displayed, only bootstrap values <100% are indicated) or triOGs (present in >24 species), a distance tree based on a superalignment of pan-triOGs, and a Consense supertree based on phylogenetic distance trees of pan-triOGs. b) Gene content tree. Bio-NJ distance tree based on the InparanOG gene content distance between two species (see "Methods" and supp. mat.). Like the other gene content trees, this tree indicates convergence in gene content of species with similar life styles.

3.4 Gene content trees have a phenotypic bias

Compared to the other phylogenomic methods, the gene content trees perform relatively poorly at recovering the required target nodes: on average, they only recover 10.38 nodes. Several numbers stand out in Fig. 2. While almost all the other trees group the Hymenomycetes, (Sce, Smi, Spa) and (Ago, Kla, Kwa, Skl) together, none of the gene content trees recover these nodes. The distance based gene content trees also fail to retrieve the Ascomycota as a monophyletic group, although this proves to be a problem for most superdistance trees as well. Interestingly, we find that part of the explanation for these biases can be found in the lifestyle of the Fungi (Fig. 3b). Although *Sch. pombe* shares relatively many genes with the Basidiomycota (supp. mat.), and might thus be expected to cluster at the root of the Ascomycota, the main dichotomy we find within the gene content tree of the Fungi is between the yeasts

on the one hand, and the filamentous fungi on the other. The dimorphic fungi, *Sch. pombe*, *Y. lipolytica* and in some cases *Can. albicans* as well, are more or less placed in between these two branches. The filamentous *P. chrysosporium* is drawn closer to the filamentous Euascomycetes within the Basidiomycota, breaking up the Hymenomycetes, and leaving the dimorphic *Cry. neoformans* and *U. maydis* as the more derived Basidiomycota in most trees. The filamentous *Ash. gossypii* stays close to its relatives, *K. lactis* and *K. waltii*, but the (Ago, Kla, Kwa, Skl) branch is never intact in the gene content trees: *Sac. kluyveri* is often at the root of this cluster. This may be a remnant genome size effect, as *Sac. kluyveri* is a very incompletely sequenced genome. To investigate the effect of the small outgroup *E. cuniculi* on the position of *Sac. kluyveri*, we removed *E. cuniculi* from the data set and recomposed the BioNJ distance tree based on the InparanOG gene content distance (Fig. 3b). The position of *Sac. kluyveri* did not alter (not shown).

This strong phenotypic effect does not explain the inability of gene conent to reproduce the target branching order in the Saccharomyces *sensu stricto* branch. In part, this may be explained by the fact that the genome sequences of *Sac. bayanus*, *Sac. kudriavzevii* and *Sac. mikatae* only covered 85 to 95% (Cliften, et al., 2003). Another issue that may specifically hinder the correct inference of the Saccharomyces *sensu stricto* branching order are differential gene losses following the complete genome duplication or alloploid genome fusion in these species (Langkjaer, et al., 2003; Scannell, et al., 2006; Wolfe and Shields, 1997). Due to the large number of redundant genes that resulted from this event, and the differential processes of gene loss that followed in the descendant lineages, a patchwork of overlapping gene repertoires will have been the result. Although such gene losses should not be in conflict with the evolutionary signal, it may be part of the reason that the gene content approaches were confounded, resulting in the deviations from the target phylogeny within the Saccharomyces *sensu stricto* clade.

3.5 Suggestions for the unresolved nodes in the fungal taxonomy

The target nodes we selected from the literature were recovered in most of our phylogenomic trees (Fig. 2). This high recovery rate supports our perhaps subjective golden standard phylogeny. In addition we were faced with three nodes that remained ambiguous in our review of the literature (supp. mat.): the internal resolution of the (Ago, Kla, Kwa, Skl) partition; the most primitive clade in the Euascomycetes; and the most primitive clade in the Ascomycota (bold lines in Fig. 2). In Table 2, we have scored the support for each of the possible branching orders in these unresolved nodes over the four phylogenomic approaches. Based on our phylogenomic data, we can make some careful conclusions about the issues that remained unresolved in the fungal phylogeny thus far.

In virtually all phylogenomic trees reconstructed in the current research, *Ash. gossypii* and *K. lactis* are sister species in the (Ago, Kla, Kwa, Skl) branch. In fact the literature references that reject this hypothesis do so with low support (Diezmann, et al., 2004; Kurtzman, 2003), while the references that support it present well supported nodes (Jeffroy, et al., 2006; Kuramae, et al., 2006; Tehler, et al., 2003). All the phylogenomic approaches support a clustering of *K. waltii* and *Sac. kluyveri*, except for the gene content trees. This suggests that the correct phylogeny is ((Ago, Kla), (Kwa, Skl)), as we also found in the high-scoring phylogenomic tree in Fig. 3a.

Table 2. Support among the trees in each of the phylogenomic approaches for the different possible branchings in the unresolved nodes of the fungal taxonomy.

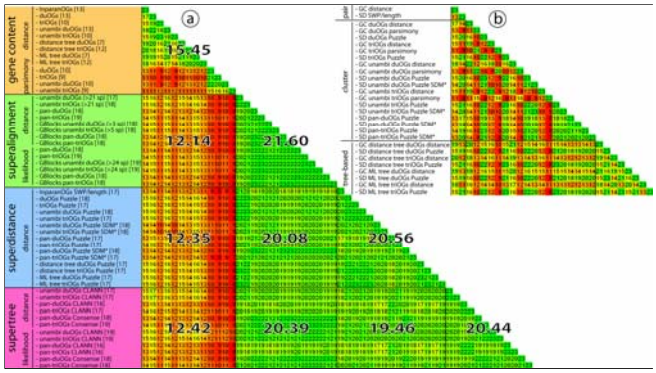
	Ago, Kla	Ago, Kwa	Ago, Skl	Kla, Kwa	Kla, Skl	Kwa, Skl	Ago, Kla, Kwa	Ago, Kla, Skl	Ago, Kwa, Skl	Kla, Kwa, Skl	(Sord,(Euro,Sno))	(Euro,(Sord,Sno))	(Sno,(Sord,Euro))	(Hemi,(Eu,Arch))	(Eu,(Hemi,Arch))	(Arch,(Hemi,Eu))
Gene content (13)	10	0	0	1	0	0	8	0	0	0	4	5	4	0	11	0
Superalignment (14)	14	0	0	0	0	14	0	0	0	0	14	0	0	0	0	12
Superdistance (15)	14	0	0	0	0	15	0	0	0	1	13	0	2	2	0	1
Supertree (12)	12	0	0	0	0	12	0	0	0	0	12	0	0	4	2	4

Our phylogenomic trees are also quite consistent regarding which clade should be placed at an ancestral position in the Euascomycetes (blue bold line in Fig. 2). Except for two of the superdistance trees, all sequence-based trees agree that *Sta. nodorum* groups with the Eurotiomycetes, and the Sordariomycetes are ancestral (Table 2). This is largely supported by the literature (Lopandic, et al., 2005; Robbertse, et al., 2006; Tehler, et al., 2003), while the only contradictory references contain other Pleosporales or Dothideomycetes, but not the species *Sta. nodorum* itself. Strikingly, the *Sta. nodorum* node is the single ill-supported node in a recent analysis of Ascomycota (Robbertse, et al., 2006).

The solution to the third unresolved issue, that of which is the most primitive of the three Ascomycotal clades (black bold line in Fig. 2), is less evident than the two above. The initial hypothesis was that *Sch. pombe* would be the first to branch off the Ascomycotal lineage (hence the name Archiascomycetes), which is also supported by most, but not all, literature references (supp. mat.). In all but two of the gene content trees the Euascomycetes are the most primitive Ascomycota, even though *Sch. pombe* clearly shares more genes with the Basidiomycota than do the other Ascomycota (supp. mat.). Conversely, the superalignment trees confidently provide the Archiascomycetes with this label, and the superdistance trees and the supertrees are inconclusive. As the superalignment trees have correctly recovered most of the other nodes as well, we conclude that their placement of the Archiascomycetes as the most primitively branching ascomycotic clade is the most reliable. Thus, the topology depicted in Fig. 3a is our final suggestion for the fungal phylogeny.

4 CONCLUDING REMARKS

We have systematically compared four phylogenomic approaches in parallel with three orthology definitions that define OGs at different levels of resolution. Using various algorithms and tree building methods, we composed a total of 54 fully automated phylogenomic trees. The main dichotomy in the topologies of the reconstructed trees is that between trees reconstructed using a sequence-based method, and trees reconstructed using gene content data (Fig. 4). The phylogenomic trees that best reproduced the target phylogeny can be found among the superalignment trees and the supertrees, using either unambiguous cluster OGs or pan-triOGs. However, although these approaches can yield trees that are completely consistent with the current opinions on the fungal phylogeny, they are not a guarantee for a successful phylogenomic tree.



For example, the CLANN supertrees based on pan-duOGs still only retrieved 16 of the 19 target nodes.

Fig. 4. Similarity between the phylogenomic trees composed in this research, ordered based on a) the phylogenomic approach and b) the orthology approach. As superalignment trees and supertrees can not use pairwise or tree-based orthology, these approaches are excluded from figure b. The small numbers in the matrices are the number of partitions shared between each pair of trees. These numbers are color coded: green (max. 23) indicates many shared partitions, red indicates few shared partitions in the tree. The large numbers are the average number of shared partitions between all trees in the four main phylogenomic approaches.

Gene content trees recover relatively few of the target nodes. This is at least partly due to convergence in the gene repertoires of Fungi with comparable phenotypes: the evolutionary and phenotypic signals are combined in one tree (Snel, et al., 1999). For example, we observe that the filamentous Euascomycetes and *P. chrysosporium* are drawn closer together, breaking the generally accepted topology of both the Ascomycota and the Basidiomycota (e.g. Fig. 2). While prokaryotes from different lineages have previously been shown to assume convergent gene repertoires in comparable ecological niches (Zomorodipour and Andersson, 1999), this is the first time (to our knowledge) that a parallel between convergence in gene content and in phenotype has been shown in Eukaryotes, to the extent that it affects a gene content phylogeny.

This research strongly supports the fungal phylogeny as displayed in Fig. 3a. The node that was recovered by the fewest phylogenomic trees is the basal position of the Archiascomycetes, represented by *Sch. pombe* here, within the Ascomycota. All other nodes are supported by many of the trees (see Fig. 2 and Table 2). Although most of these branches are supported by recent literature (Table 1 in supp. mat.), this research helped provide support for those cases that were inconclusive (Table 2 and Table 2 in supp. mat.). What is striking in our phylogenetic findings is that that several of the fungal groups presented in the Genbank Taxonomy Database (Wheeler, et al., 2002) should actually be adjusted. For example, *Candida*, *Kluyveromyces*, *Saccharomyces* and the Saccharomycetaceae remain mentioned as clades, while their members should be regrouped (see also (Diezmann, et al., 2004; Kurtzman, 1998; Kurtzman, 2003; Lopandic, et al., 2005; Prillinger, et al., 2002; Tehler, et al., 2003)).

Our phylogenomic trees of the Fungi reproduced many of the clades in accordance with the current taxonomic views. At least for the Fungi, we confirm a number of standard practices in the current phylogenomics field, albeit it with small differences relative to the less well-established approaches such as supertrees. A recent superalignment tree (Ciccarelli, et al., 2006) has been criticised as being a "tree of one percent" of the genome (Dagan and Martin, 2006). In the current study, we show that methods that are restrictive in selecting genes often create a phylogeny that is close to the golden standard. Apparently, this selection procedure is necessary to filter out the noise caused by evolutionary processes like gene duplication and gene loss, even in the absence of horizontal transfer (Andersson, 2005). Complete genomes allow us to do this automatically and still retain enough genes to construct a reliable phylogeny. Our results indicate that a (1) maximum likelihood (2) superalignment tree based on (3) selected well aligned positions of (4) unambiguous cluster OGs, automatically derived at the level of resolution most suitable for the group of species considered, will yield a respectable tree. Maximum likelihood (1), because we find that distance trees may have trouble with the outgroup we used in this study; superalignment (2), because on average, this phylogenomic approach recovers the most target nodes; unambiguously aligned positions (3), because this enables the inclusion of more high quality data; and finally unambiguous cluster OGs derived at the level of the taxon of interest (4), because this ensures that you have the highest resolution possible.

REFERENCES

(http://www.broad.mit.edu) Broad Institute of MIT and Harvard.

(http://www.jgi.doe.gov) DOE Joint Genome Institute.

Andersson, J.O. (2005) Lateral gene transfer in eukaryotes, *Cell Mol Life Sci*, **62**, 1182-1197.

Berbee, M.L., Carmean, D.A. and Winka, K. (2000) Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough? *Mol Phylogenet Evol*, **17**, 337-344.

Bininda-Emonds, O.R.P. (2004) The evolution of supertrees, *Trends in Ecology & Evolution*, **19**, 315-322.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol Biol Evol*, **17**, 540-552.

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.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A. and Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting, *Science*, **301**, 71-76.

Creevey, C.J. and McInerney, J.O. (2005) Clann: investigating phylogenetic information through supertree analyses, *Bioinformatics*, **21**, 390-392.

Criscuolo, A., Berry, V., Douzery, E.J.P. and Gascuel, O. (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics, *Syst Biol*, In press.

Dagan, T. and Martin, W. (2006) The tree of one percent, *Genome Biol*, **7**, 118.

Daubin, V., Gouy, M. and Perriere, G. (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history, *Genome Res*, **12**, 1080-1090.

Daubin, V., Moran, N.A. and Ochman, H. (2003) Phylogenetics and the cohesion of bacterial genomes, *Science*, **301**, 829-832.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., et al. (2005) The genome sequence of the rice blast fungus Magnaporthe grisea, *Nature*, **434**, 980-986.

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

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., et al. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome, *Science*, **304**, 304-307.

Diezmann, S., Cox, C.J., Schonian, G., Vilgalys, R.J. and Mitchell, T.G. (2004) Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis, *J Clin Microbiol*, **42**, 5624-5635.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree, *Science*, **284**, 2124-2129.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., et al. (2004) Genome evolution in yeasts, *Nature*, **430**, 35-44.

Dutilh, B.E., Huynen, M.A., Bruno, W.J. and Snel, B. (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise, *J Mol Evol*, **58**, 527-539.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792-1797.

Farris, R.J. (1977) Phylogenetic analysis under Dollo's law, *Syst Zool*, **26**, 77-88.

Fell, J.W., Boekhout, T., Fonseca, A., Scorzetti, G. and Statzell-Tallman, A. (2000) Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis, *Int J Syst Evol Microbiol*, **50 Pt 3**, 1351-1371.

Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2), *Cladistics*, **5**, 164-166.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins, *Syst Zool*, **19**, 99-113.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., et al. (2003) The genome sequence of the filamentous fungus Neurospora crassa, *Nature*, **422**, 859-868.

Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.J., Wortman, J.R., Batzoglou, S., Lee, S.I., Basturkmen, M., Spevak, C.C., Clutterbuck, J., et al. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae, *Nature*, **438**, 1105-1115.

Ge, F., Wang, L.S. and Kim, J. (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer, *PLoS Biol*, **3**, e316.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996) Life with 6000 genes, *Science*, **274**, 546-&.

Guarro, J., GeneJ and Stchigel, A.M. (1999) Developments in fungal taxonomy, *Clin Microbiol Rev*, **12**, 454-500.

Hillis, D.M., Huelsenbeck, J.P. and Cunningham, C.W. (1994) Application and accuracy of molecular phylogenies, *Science*, **264**, 671-677.

Huynen, M.A., Snel, B. and Bork, P. (2001) Inversions and the dynamics of eukaryotic gene order, *Trends Genet*, **17**, 304-306.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., et al. (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny, *Nature*, **443**, 818-822.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006) Phylogenomics: the beginning of incongruence? *Trends Genet*, **22**, 225-231.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., et al. (2004) The diploid genome sequence of Candida albicans, *Proc Natl Acad Sci U S A*, **101**, 7329-7334.

Kamper, J., Kahmann, R., Bolker, M., Ma, L.J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Muller, O., et al. (2006) Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis, *Nature*, **444**, 97-101.

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.

Kellis, M., Birren, B.W. and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae, *Nature*, **428**, 617-624.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature*, **423**, 241-254.

Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies, *Trends Genet*, **18**, 158-162.

Kouvelis, V.N., Ghikas, D.V. and Typas, M.A. (2004) The analysis of the complete mitochondrial genome of Lecanicillium muscarium (synonym Verticillium lecanii) suggests a minimum common gene organization in mtDNAs of Sordariomycetes: phylogenetic implications, *Fungal Genet Biol*, **41**, 930-940.

Kunin, V., Goldovsky, L., Darzentas, N. and Ouzounis, C.A. (2005) The net of life: reconstructing the microbial phylogenetic network, *Genome Res*, **15**, 954-959.

Kuramae, E., Robert, V., Snel, B., Weiss, M. and Boekhout, T. (2006) Phylogenomics reveal a robust fungal tree of life, *FEMS Yeast Res*, In press.

Kurtzman, C.P. (1998) Discussion of teleomorphic and anamorphic ascomycetous yeasts and a key to genera. In Kurtzman, C.P. and Fell, J.W. (eds), *The yeasts, a taxonomic study*. Elsevier, Amsterdam, The Netherlands, 111-121.

Kurtzman, C.P. (2003) Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora, *FEMS Yeast Res*, **4**, 233-245.

Langkjaer, R.B., Cliften, P.F., Johnston, M. and Piskur, J. (2003) Yeast genome duplication was followed by asynchronous differentiation of duplicated genes, *Nature*, **421**, 848-852.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., et al. (2005) The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans, *Science*, **307**, 1321-1324.

Lopandic, K., Molnar, O., Suzuki, M., Pinsker, W. and Prillinger, H. (2005) Estimation of Phylogenetic relationships within the Ascomycota on the basis of 18S rDNA sequences and chemotaxonomy, *Mycol Progress*, **4**, 205-214.

Lutzoni, F., Kauff, F., Cox, C.J., McLaughlin, D., Celio, G., Dentinger, B., Padamsee, M., Hibbett, D., James, T.Y., Baloch, E., et al. (2004) Assembling the fungal tree of life: Progress, classification and evolution of subcellular traits, *American Journal of Botany*, **91**, 1446-1480.

Martinez, D., Larrondo, L.F., Putnam, N., Gelpke, M.D., Huang, K., Chapman, J., Helfenbein, K.G., Ramaiya, P., Detter, J.C., Larimer, F., et al. (2004) Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78, *Nat Biotechnol*, **22**, 695-700.

Medina, M. (2005) Genomes, phylogeny, and evolutionary systems biology, *Proc Natl Acad Sci U S A*, **102 Suppl 1**, 6630-6635.

Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., Berriman, M., Abe, K., Archer, D.B., Bermejo, C., et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus, *Nature*, **438**, 1151-1156.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W. and Casane, D. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments, *Mol Biol Evol*, **21**, 1740-1752.

Prillinger, H., Lopandic, K., Schweigkofler, W., Deak, R., Aarts, H.J.M., Bauer, R., Sterflinger, K., Kraus, G.F. and Maraz, A. (2002) Phylogeny and systematics of the fungi with special reference to the Ascomycota and Basidiomycota, *Fungal Allergy and Pathogenicity*, **81**, 207-295.

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

Robbertse, B., Reeves, J.B., Schoch, C.L. and Spatafora, J.W. (2006) A phylogenomic analysis of the Ascomycota, *Fungal Genet Biol*, **43**, 715-725.

Rokas, A., Williams, B.L., King, N. and Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature*, **425**, 798-804.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. and Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts, *Nature*, **440**, 341-345.

Scorzetti, G., Fell, J.W., Fonseca, A. and Statzell-Tallman, A. (2002) Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions, *FEMS Yeast Res*, **2**, 495-517.

Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content, *Nat Genet*, **21**, 108-110.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

Tehler, A., Little, D.P. and Farris, J.S. (2003) The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, Fungi, *Mycol Res*, **107**, 901-916.

Teichmann, S.A. and Mitchison, G. (1999) Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol*, **49**, 98-107.

Thomarat, F., Vivares, C.P. and Gouy, M. (2004) Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes, *J Mol Evol*, **59**, 780-791.

van der Heijden, R.T.J.M., Snel, B., van Noort, V. and Huynen, M.A. (submitted) Orthology prediction at scalable resolution through automated analysis of phylogenetic trees, *BMC Bioinformatics*.

Vivares, C.P., Gouy, M., Thomarat, F. and Metenier, G. (2002) Functional and evolutionary analysis of a eukaryotic parasitic genome, *Curr Opin Microbiol*, **5**, 499-505.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. (2002) Database resources of the National Center for Biotechnology Information: 2002 update, *Nucleic Acids Res*, **30**, 13-16.

Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, **387**, 708-713.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. (2002) The genome sequence of Schizosaccharomyces pombe, *Nature*, **415**, 871-880.

Zomorodipour, A. and Andersson, S.G. (1999) Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis, *FEBS Lett*, **452**, 11-15.