Université de Montréal

A phylogenomics approach to resolving fungal evolution, and phylogenetic method development

par

Yu Liu

Département de Biochimie
Faculté de Médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Doctorat
en Bio-informatique

December, 2009

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

A phylogenomics approach to resolving fungal evolution, and phylogenetic method development

Présentée par :

Yu Liu

a été évaluée par un jury composé des personnes suivantes :

Nicolas Lartillot, président-rapporteur

B Franz Lang, directeur de recherche

Miklós Csűrös, membre du jury

Christian Blouin, examinateur externe

Gertraud Burger, représentant du doyen de la FES

Résumé

Bien que les champignons soient régulièrement utilisés comme modèle d'étude des systèmes eucaryotes, leurs relations phylogénétiques soulèvent encore des questions controversées. Parmi celles-ci, la classification des zygomycètes reste inconsistante. Ils sont potentiellement paraphylétiques, i.e. regroupent de lignées fongiques non directement affiliées. La position phylogénétique du genre *Schizosaccharomyces* est aussi controversée: appartient-il aux Taphrinomycotina (précédemment connus comme archiascomycetes) comme prédit par l'analyse de gènes nucléaires, ou est-il plutôt relié aux Saccharomycotina (levures bourgeonnantes) tel que le suggère la phylogénie mitochondriale? Une autre question concerne la position phylogénétique des nucléariides, un groupe d'eucaryotes amiboïdes que l'on suppose étroitement relié aux champignons. Des analyses multi-gènes réalisées antérieurement n'ont pu conclure, étant donné le choix d'un nombre réduit de taxons et l'utilisation de six gènes nucléaires seulement.

Nous avons abordé ces questions par le biais d'inférences phylogénétiques et tests statistiques appliqués à des assemblages de données phylogénomiques nucléaires et mitochondriales. D'après nos résultats, les zygomycètes sont paraphylétiques (Chapitre 2) bien que le signal phylogénétique issu du jeu de données mitochondriales disponibles est insuffisant pour résoudre l'ordre de cet embranchement avec une confiance statistique significative. Dans le Chapitre 3, nous montrons à l'aide d'un jeu de données nucléaires important (plus de cent protéines) et avec supports statistiques concluants, que le genre *Schizosaccharomyces* appartient aux Taphrinomycotina. De plus, nous démontrons que le regroupement conflictuel des *Schizosaccharomyces* avec les Saccharomycotina, venant des données mitochondriales, est le résultat d'un type d'erreur phylogénétique connu:

l'attraction des longues branches (ALB), un artéfact menant au regroupement d'espèces dont le taux d'évolution rapide n'est pas représentatif de leur véritable position dans l'arbre phylogénétique. Dans le Chapitre 4, en utilisant encore un important jeu de données nucléaires, nous démontrons avec support statistique significatif que les nucleariides constituent le groupe lié de plus près aux champignons. Nous confirmons aussi la paraphylie des zygomycètes traditionnels tel que suggéré précédemment, avec support statistique significatif, bien que ne pouvant placer tous les membres du groupe avec confiance. Nos résultats remettent en cause des aspects d'une récente reclassification taxonomique des zygomycètes et de leurs voisins, les chytridiomycètes.

Contrer ou minimiser les artéfacts phylogénétiques telle l'attraction des longues branches (ALB) constitue une question récurrente majeure. Dans ce sens, nous avons développé une nouvelle méthode (Chapitre 5) qui identifie et élimine dans une séquence les sites présentant une grande variation du taux d'évolution (sites fortement hétérotaches - sites HH); ces sites sont connus comme contribuant significativement au phénomène d'ALB. Notre méthode est basée sur un test de rapport de vraisemblance (likelihood ratio test, LRT). Deux jeux de données publiés précédemment sont utilisés pour démontrer que le retrait graduel des sites HH chez les espèces à évolution accélérée (sensibles à l'ALB) augmente significativement le support pour la topologie « vraie » attendue, et ce, de façon plus efficace comparée à d'autres méthodes publiées de retrait de sites de séquences. Néanmoins, et de façon générale, la manipulation de données préalable à l'analyse est loin d’être idéale. Les développements futurs devront viser l'intégration de l'identification et la pondération des sites HH au processus d'inférence phylogénétique lui-même.

Mots-clés : phylogénomique, Taphrinomycotina, zygomycètes, attraction des longues branches, mitochondrial, nucleariides, hétérotache, likelihood ratio test.

Abstract

Despite the popularity of fungi as eukaryotic model systems, several questions on their phylogenetic relationships continue to be controversial. These include the classification of zygomycetes that are potentially paraphyletic, i.e. a combination of several not directly related fungal lineages. The phylogenetic position of *Schizosaccharomyces* species has also been controversial: do they belong to Taphrinomycotina (previously known as archiascomycetes) as predicted by analyses with nuclear genes, or are they instead related to Saccharomycotina (budding yeast) as in mitochondrial phylogenies? Another question concerns the precise phylogenetic position of nucleariids, a group of amoeboid eukaryotes that are believed to be close relatives of Fungi. Previously conducted multi-gene analyses have been inconclusive, because of limited taxon sampling and the use of only six nuclear genes.

We have addressed these issues by assembling phylogenomic nuclear and mitochondrial datasets for phylogenetic inference and statistical testing. According to our results zygomycetes appear to be paraphyletic (Chapter 2), but the phylogenetic signal in the available mitochondrial dataset is insufficient for resolving their branching order with statistical confidence. In Chapter 3 we show with a large nuclear dataset (more than 100 proteins) and conclusive supports that *Schizosaccharomyces* species are part of Taphrinomycotina. We further demonstrate that the conflicting grouping of *Schizosaccharomyces* with budding yeasts, obtained with mitochondrial sequences, results from a phylogenetic error known as long-branch attraction (LBA, a common artifact that leads to the regrouping of species with high evolutionary rates irrespective of their true phylogenetic positions). In Chapter 4, using again a large nuclear dataset we demonstrate with significant statistical support that nucleariids are the closest known relatives of Fungi. We also confirm paraphyly of traditional zygomycetes as previously suggested, with significant support, but without placing all members of this group with confidence. Our results question aspects of a recent taxonomical reclassification of zygomycetes and their chytridiomycete neighbors (a group of zoospore-producing Fungi).

Overcoming or minimizing phylogenetic artifacts such as LBA has been among our most recurring questions. We have therefore developed a new method (Chapter 5) that identifies and eliminates sequence sites with highly uneven evolutionary rates (highly heterotachous sites, or HH sites) that are known to contribute significantly to LBA. Our method is based on a likelihood ratio test (LRT). Two previously published datasets are used to demonstrate that gradual removal of HH sites in fast-evolving species (suspected for LBA) significantly increases the support for the expected ‘true’ topology, in a more effective way than comparable, published methods of sequence site removal. Yet in general, data manipulation prior to analysis is far from ideal. Future development should aim at integration of HH site identification and weighting into the phylogenetic inference process itself.

Keywords : phylogenomics, Taphrinomycotina, Zygomycota, long-branch attraction, mitochondrial, nucleariids, heterotachous, likelihood ratio test

Table of contents

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

To my parents and family

Acknowledgements

My first thank you goes to my supervisor Dr B. Franz Lang. First, I would like to thank him for giving me the opportunity to work with him. His approach to supervision has forced me to be independent and to work my way through problems, and for that I am especially grateful. I also appreciate his patience and teaching, like sitting through all the painful practice talks and improving my writing skills.

A huge thank you goes further to Dr Herve Philippe for his most useful programs, papers, datasets and critical comments to my work. Basically, I learned phylogenetics from his papers and talks. I also thank Dr Gertraud Burger for her help in many aspects; her comments and suggestions during group meetings are greatly appreciated. I also want to thank Dr Nicolas Lartillot for the development of PhyloBayes that implements the CAT model, and for his help to use it.

I would also like to thank all the past and current members of our group and the Robert Cedergren Centre, including Jessica Leigh, Nicolas Rodrigue, and Rachel Bevan for their help in writing papers; Henner Brinkmann for his help in phylenetics analysis; Amy Hauth, Lise Forget, Veronique Marie, David To, Allan Sun, Eric Wang for advice and suggestions in programming; Yan Zhou and Yaoqing Shen for the wonderful discussions about phylogeny inference and molecular biology.

I thank my thesis committee members, Dr Nicolas Lartillot, Dr Miklós Csűrös, and Dr Christian Blouin, for their insightful comments to improve this thesis. Elaine Meunier deserves my special thanks. Her kind help makes things much easier for a student who cannot speak French in a francophone university.

Last but not least, I would like to thank my parents and family for supporting me spiritually throughout my life.

Chapter 1 Introduction

Fungi are widely used as model systems in molecular and cellular biology. In this context, it is crucial to understand and evaluate their respective evolutionary relationships and how their genes and genomes change over time. Yet, after decades of research, numerous questions related to fungal evolution remain unresolved. In this thesis, we will focus on three topics of general interest:

- are the Taphrinomycotina (including *Schizosaccharomyces* species) monophyletic or paraphyletic (for a current view of fungal taxonomy see Table 1)?
- what are the relationship among the major fungal groups, especially the less well known zygomycetes and chytrids?
- what are the exact phylogenetic positions of protists that are believed to diverge close to the metazoan-fungal boundary, such as Nucleariida, *Capsaspora*, *Amoebidium* and *Sphaeroforma* (for taxonomic details see Table 2)?

To do so, we analyzed nuclear and mitochondrial genomic datasets with various phylogenetic methods. We also developed a new method to improve the accuracy of phylogenetic inference.

In the first chapter of this dissertation we will present an overview of current fungal taxonomy and the status of molecular phylogenetic inference. The models and methods used in phylogenetic analyses are also described in this chapter, along with a discussion of analytical challenges that are related to common phylogenetic artifacts. Chapters two to five present our results in the format of journal publications (three published and one manuscript). They aim essentially at resolving issues in fungal and opisthokont evolution.

The first publication (Chapter 2) presents comparative mitochondrial genomic analyses of zygomycetes and tests the monophyly of zygomycetes. We find that the mitochondrial dataset alone is insufficient to resolve with confidence whether Zygomycota is a monophyletic or paraphyletic group, and members of this group as well as Fungi in general evolve at a wide range of evolutionary rates. These rate differences may cause Long-Branch Attraction artifact (LBA, causing grouping of fast-evolving lineages irrespective of their true evolutionary relationships, sometimes even with strong statistical support), a theme that is addressed in detail in the following publications. The second publication (Chapter 3) presents phylogenomic analyses that aim to the issue of Taphrinomycotina. Our results suggest that Taphrinomycotina is a monophyletic group, a sister group of Saccharomycotina plus Pezizomycotina, and that a LBA artifact plagues the analyses of mitochondrial data and leads to a paraphyletic Taphrinomycotina. In the third publication (Chapter 4), our analyses with both nuclear and mitochondrial genes confirm that nucleariids are the closest unicellular relatives of Fungi; that *Capsaspora*, *Amoebidium* plus *Sphaeroforma* form a monophyletic sister group of Metazoa plus Choanoflagellata; and that Zygomycota and Chytridiomycota as defined in traditional taxonomy are most likely paraphyletic.

Table 1: Fungal Systematics (according to Hibbett et al. 2007)

Taxon	Examples of member species
Phylum Ascomycota	
Subphylum Taphrinomycotina	*Schizosaccharomyces pombe*
Subphylum Saccharomycotina	*Saccharomyces cerevisiae*
Subphylum Pezizomycotina	*Neurospora crassa*
Phylum Basidiomycota	
Subphylum Urediniomycotina	*Puccinia graminis*
Subphylum Ustilaginomycotina	*Ustilago maydis*
Subphylum Hymenomycotina	*Cryptococcus neoformans*
Phylum Chytridiomycota[1]	*Spizellomyces punctatus*
Phylum Neocallimastigomycota[1]	*Neocallimastix frontalis*
Phylum Blastocladiomycota[1]	*Allomyces macrogynus*
Phylum Glomeromycota[2]	*Glomus intraradices*
Subphylum* Mucoromycotina[2]	*Rhizopus oryzae*
Subphylum* Kickxellomycotina[2]	*Smittium culisetae*
Subphylum* Zoopagomycotina[2]	*Zoophagus insidians*
Subphylum* Entomophthoromycotina[2]	*Conidiobolus coronatus*

* currently not assigned to a phylum

[1] Chytridiomycota in traditional taxonomy

[2] Zygomycota in traditional taxonomy

Table 2: Unicellular protists that are believed to diverge close to Fungi/Metazoa

Taxon	Examples of member species	Reference
Choanoflagellida	*Monosiga brevicollis*	(King et al. 2008)
Corallochytrium	*Corallochytrium limacisporum*	(Sumathi et al. 2006)
Eccrinales	*Alacrinella limnoriae*	(Cafaro 2005)
Ichthyosporea		
Capsaspora	*Capsaspora owczarzaki*	(Ruiz-Trillo et al. 2008)
Dermocystida	*Dermocystidium salmonis*	(Marshall et al. 2008)
Ichthyophonida		
Amoebidiaceae	*Amoebidium parasiticum*	(Ruiz-Trillo et al. 2008)
Sphaeroforma	*Sphaeroforma arctica*	(Ruiz-Trillo et al. 2006)
Microsporida	*Encephalitozoon cuniculi*	(Keeling 2009)
Ministeria	*Ministeria vibrans*	(Shalchian-Tabrizi et al. 2008)
Nucleariidae	*Nuclearia simplex*	(Steenkamp et al. 2006)
Rozellida	*Rozella allomycis*	(Lara et al. 2009)

The fifth chapter deals with method development, addressing limitations of current approaches in phylogenetic analysis due to model violations. Here we describe a new method that improves the information/background noise ratio in datasets by progressive elimination of positions that likely contribute to LBA. The method is based on a Likelihood Ratio Test (LRT; a comparison of likelihood values under two models), which is used to identify and eliminate sequence positions that contain little if any phylogenetic signal. Sequence elimination occurs specifically in fast evolving species (or groups) as they are most affected by LBA. Two previously published datasets are used to demonstrate the potential of this method. The result shows that it can effectively overcome LBA. Finally, in the sixth chapter we summarize the most important findings of this dissertation, compare them with previous work, and comment on future orientations of our research.

1. Fungal classification

Based on morphology, ultra-structural characteristics, and lifestyle, Fungi include more than one million extremely diverse species, a major challenge for morphology-based fungal taxonomy (Hawksworth 2001). Today, sequence-based, molecular taxonomy clearly defines Fungi as a monophyletic group, further revealing their close relationship to animals (Metazoa) rather than to plants as previously believed. The following section will provide an overview of the current molecular taxonomy of Fungi.

1.1 Relationship of Fungi with other eukaryotic groups.

Eukaryotes include organisms with an almost inconceivable morphological diversity. Not surprisingly, their classification has changed dramatically over the decades, from four major groups or kingdoms (plants, animals, fungi, and protists) to the current system of six “super-groups”: Amoebozoa, Chromalveolata, Excavata, Opisthokonta, Plantae, Rhizaria (Parfrey et al. 2006; Rodriguez-Ezpeleta et al. 2007a; Baldauf 2008; Yoon et al. 2008). The supergroup “Plantae” contains three lineages with primary plastids: green algae (including land plants), rhodophytes, and glaucophytes (the term primary plastids refers to the original endosymbiotic event of a cyanobacterium with a eukaryote that gave rise to plastids; for more details see (Rodriguez-Ezpeleta et al. 2005; Kim and Graham 2008)). Fungi have traditionally been considered a subgroup of plants, but phylogenetic analyses of both nuclear small subunit rRNA (SSU-rRNA) and protein-coding genes clearly reject this association. These analyses provide instead significant support for a sister group relationship of Fungi with Metazoa (Opistokonta), now also including a number protists (Baldauf et al. 2000; Moreira, Le Guyader, and Philippe 2000; Lang et al. 2002b; Cavalier-Smith 2004; Parfrey et al. 2006; Yoon et al. 2008). In addition, some groups previously classified as Fungi (such as Myxomycota, Dictyosteliomycota, and Oomycota) are now excluded from the kingdom (Gunderson et al. 1987; Paquin et al. 1997).

Protists are a taxonomically inconsistent group uniting diverse eukaryotic organisms that have not been associated with animals, fungi or plants. Based on morphologic markers, protists were traditionally subdivided into animal-like, plant-like, and fungus-like groups, a classification that does not reflect the protists' true evolutionary relationships (Cavalier-Smith and Chao 2003b; Adl et al. 2005). Molecular phylogenetic analyses have revealed that some protists are indeed Fungi (e.g., *Pneumocystis carinii* (Edman et al. 1988a)), and that others such as Microsporidia (Keeling 2003) and Nucleariida (Steenkamp, Wright, and Baldauf 2006) are related to them. Yet, the exact phylogenetic position of Microsporidia is uncertain due to their extremely fast evolutionary rate, (Hibbett et al. 2007) and that Nucleariida are likely a sister group of Fungi - awaiting confirmation by additional, statistically more compelling analyses (Steenkamp, Wright, and Baldauf 2006). The phylogenetic positions of other potential opisthokont relatives (e.g., Apusozoa and potentially Malawimonadozoa) remains currently uncertain, even with phylogenomic datasets (Philippe 2000; Parfrey et al. 2006) (Lang and Philippe, unpublished results).

1.2 Fungal subgroups.

Based on their sexual reproductive structures, Fungi have been traditionally divided into Ascomycota, Basidiomycota, Zygomycota, and Chytridiomycota (Taylor et al. 2004; McLaughlin et al. 2009). However, the classification has changed dramatically in recent years (Seif et al. 2005; James et al. 2006a; Liu, Hodson, and Hall 2006; Spatafora et al. 2006; Hibbett 2007), especially for the Zygomycota and Chytridiomycota. In the most recent classification, Chytridiomycota remains a phylum but in a restricted sense, now only including Chytridiomycetes and Monoblepharidomycetes. Other traditional members such as Blastocladiomycota and Neocallimastigales are elevated to separate phyla, and the phylum Zygomycota is completely abandoned. Its member are divided into the phylum Glomeromycota plus four subphyla *incertae sedis* (not assigned to any phylum): Mucoromycotina, Kickxellomycotina, Zoopagomycotina and Entomophthoromycotina. To some extent, this classification is consistent with phylogenetic analyses based on single rRNA and protein-coding genes, on combinations of few genes, and most reliably, on nuclear and mitochondrial multi-gene datasets (e.g., (Edman et al. 1988a; Lang et al. 2002a; Bullerwell, Forget, and Lang 2003b; Leigh et al. 2003; Thomarat, Vivares, and Gouy 2004; Fitzpatrick et al. 2006; James et al. 2006a; Liu, Hodson, and Hall 2006;

Hibbett et al. 2007)). Yet, only Ascomycota and Basidiomycota are clearly monophyletic sister clades, and several higher-order relationships among other fungal lineages remain uncertain. The lack of phylogenetic comprehension renders naming of new and abandoning of previously established taxonomic groups challenging, if not controversial (e.g., Blastocladiomycota and Glomeromycota; Zygomycota; Hibbett et al. 2007).

2. Unresolved issues in fungal phylogeny

2.1 Monophyly of Taphrinomycotina and relationships among major fungal lineages

Molecular taxonomies based on SSU-rRNA sequences divide Ascomycota into three major lineages: Saccharomycotina, Pezizomycotina and Taphrinomycotina (Archiascomycota) (Nishida and Sugiyama 1993). In this analysis, Taphrinomycotina is the sister group of Saccharomycotina plus Pezizomycotina. However, significant support for this topology is lacking, and results from several multi-protein phylogenies using mitochondrial and nuclear protein sequences are incongruent (e.g., (Bullerwell et al. 2003; Taylor et al. 2004)). The reason for this incongruence appears to be a LBA attraction artifact. It is observed both, with certain nuclear (Baldauf et al. 2000) and mitochondrial datasets (Leigh et al. 2004). In both cited cases, LBA leads to an incorrect, yet statistically well-supported grouping of Saccharomycotina plus *Schizosaccharomyces* (a key member of Taphrinomycotina and widely used model system). Adding more Taphrinomycotina (*Taphrina*, *Saitoella*, *Pneumocystis* and *Neolecta*) and more sequences from this species has potential to overcome this phylogenetic artifact.

Besides unresolved issues within Ascomycota, other higher-order relationships among other fungal lineages also remain uncertain, such as zygomycetes (Keeling, Luker, and Palmer 2000; Schwarzott, Walker, and Schussler 2001; Forget et al. 2002; Tehler, Little, and Farris 2003; Leigh et al. 2004; Seif et al. 2005; Tanabe, Watanabe, and Sugiyama 2005b; James et al. 2006a; Liu, Hodson, and Hall 2006; Hibbett 2007). One reason for this uncertainty is missing sequence, in particular in the large phylogenomic datasets: complete genome sequences are available from only few lineages in zygomycetes and chytridiomycetes. Another reason is the presence of short internal branches that separate them and their sub-groups. More (and more complete) genomic data, and the development of increasingly sophisticated evolutionary models and phylogenetic algorithms that are better in extracting ‘phylogenetic signal’ are hoped to resolve these issues.

2.2 Unresolved phylogenetic relationship between protists and Fungi

Some protists such as the choanoflagellates (Choanoflagellata), the ichthyosporeans (Ichthyosporea), the nucleariids (Nucleariidae) and the genera *Capsaspora* and *Ministeria* are believed to branch close to the fungal-animal divergence. In some instances, their exact phylogenetic positions are still debated (Cavalier-Smith and Chao 2003b; Adl et al. 2005; Steenkamp, Wright, and Baldauf 2006; Carr et al. 2008; King et al. 2008; Ruiz-Trillo et al. 2008). Among them, nucleariids is of particular interest, the only group that appears to be closely related to Fungi. The few molecular phylogenies including nucleariids are based on single gene sequences (except (Steenkamp, Wright, and Baldauf 2006), see below), and come with contradicting results and unconvincing statistical support. For instance, nucleariids are the sister of Fungi in an analysis of SSU plus LSU data, yet with limited species sampling (Medina et al. 2003). In contrast, in analyses with SSU data and rich taxa sampling, nucleariids are the sister of Metazoa/Choanoflagellata/Mesomycetozoa (Zettler et al. 2001). In a more recent multi-gene analysis (including EF-1α, actin, HSP70 and α- and β-tubulin) they are the sister group of Fungi, yet despite the use of multiple gene sequences, competing tree topologies cannot be rejected with confidence (Steenkamp, Wright, and Baldauf 2006). Evidently, much larger datasets are required to resolve this question.

3. Evolutionary models for amino acid sequence change

Nucleotide and amino acid sequences are two of the principle types of molecular data used in phylogenetic analyses. The third codon positions of nucleotide sequences of protein-coding genes have higher evolutionary rates and a more pronounced compositional sequence bias (Li 1997), which may lead to incorrect phylogenetic inferences as described in the next sections. Thus, our phylogenomic datasets contain exclusively derived protein-coding sequences. Models that describe evolutionary amino acids change are an essential component of phylogenetic inference. This section summarizes the statistical basis and properties of the evolutionary models for amino acid sequences.

3.1 Markov process models

The aim of this section is to introduce the common assumptions made to model the process of molecular evolution.

One of the primary assumptions is that future evolution (at time T = t+1) is only dependent on its current state (at time T = t) and not on previous states (T < t). The processes with this property are called Markov processes in statistics. This assumption is reasonable, as mutation and substitution can only act upon the present molecules in an organism.

To reduce the complexity of evolutionary models and the computational burden, another common assumption is that sequence sites of molecules evolve independently, although this does not always hold in real sequence. Yet, studies show that models with this assumption perform reasonably well. Recently, models that relax this assumption have been developed (Galtier 2001; Penny et al. 2001), but due to their heavy computational cost they are not widely used and tested for performance.

Based on these assumptions, various global amino acid substitutions models have been developed during the last four decades (Dayhoff 1978; Gonnet, Cohen, and Benner 1992; Jones, Taylor, and Thornton 1992; Whelan and Goldman 2001; Le and Gascuel 2008). In the following section, we will describe their common principles and also explain a more recently developed model that describes site-specific change: CAT.

3.2 Instantaneous rate and probability matrices

A continuous-time Markov process is used to model amino acid changes at a given site of a protein sequence. The Markov model asserts that one protein sequence is derived from its ancestor by a series of independent substitutions. For protein sequences, the continuous-time Markov process is defined by its instantaneous 20×20 rate matrix:

$$Q = (q_{ij}), \ i,j=1,..20$$

The matrix entry q_{ij}, $j \neq i$, represents the instantaneous rate of change from amino acid i to amino acid j, independently at each site. The q_{ii} entry of the matrix is set to be the minus of the sum of all other entries in that row, representing the rate at which changes leave amino acid i and making the row sums being zero:

$$q_{ii} = -\sum_{j,j \neq i}^{N} q_{ij}$$

However, protein sequence data consist of actual amino acid characters at a given time, not the rate at which they are evolving. The quantity needed for calculations is the probability of observing a given character after time t has elapsed. Let $P_{ij}(t)$ be the probability of a site in state j after time t, given that the process started in state i at that site at time 0. Since there are 20 amino acid states, the probability $P_{ij}(t)$ can be written as a 20×20 matrix, which denoted as P(t). It is necessary to compute the probability matrix P(t) for any real evolutionary time $t \geq 0$. This is achieved using the instantaneous rate matrix Q, which is related to P(t) *via* $P(t) = e^{tQ}$. The exponential of a matrix is defined by the following power series, with I being the identity matrix:

$$e^{-tQ} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \ldots$$

In practice, this power series can be calculated numerically using eigendecomposition and other methods.

A Markov process model has three important properties: homogeneity, stationarity, and reversibility. Homogeneity means that the rate matrix Q is independent on time T, which means that the patterns of amino acid substitution remain the same during evolutionary history. A homogeneous process has an equilibrium distribution that is the distribution when time approaches infinity. Stationarity means that the process is at that equilibrium, which implies amino acids frequencies have remained the same during the course of evolution. Reversibility means that $\pi_i P_{ij}(T) = \pi_j P_{ji}(T)$ for all *i, j*, and *T*, where π_i are the frequencies of occurrence for each amino acid.

3.3 Rate heterogeneity among sites

Due to the various functional constraints on sites, evolutionary rates vary at different sites along the sequence. In other words, it is less likely that substitutions occur at positions with strong functional constraints, and amino acids at positions with few constraints are more easily substituted. A common way of modeling evolutionary rates along a sequence is by applying a Gamma distribution (a density function),

$$y = f(x|a,b) = \frac{1}{b^a \ \Gamma(a)} \ x^{a\text{-}1} e^{-\frac{x}{b}}$$

where Γ is the Gamma function, parameter *a* the shape parameter, and *b* the inverse scale parameter. The mean of a Gamma-distributed variable is a/b; the variance is a/b^2. The Gamma distribution is sufficiently general to accommodate different levels of rate heterogeneity in various datasets, and usually, $b = a$ is assumed. The shape parameter *a* determines the extent of rate heterogeneity among sites, with a small *a* representing extreme rate variation (Yang 1996). However, as pointed out correctly, "there is nothing about the Gamma distribution that makes it biologically more realistic than any other distribution … It is used because of its mathematical tractability" (Felsenstein 2004).

In practice, most applications provide a discrete Gamma distribution (Yang 1996), because a continuous likelihood calculation is computationally too demanding. Studies have shown that a discrete gamma distribution with four to eight categories provides both a good approximation and reasonable computational efficiency (Yang 1996).

3.4 CAT model

Most of models assume that sites along a sequence are independent and identically distributed (*i.i.d.*). While this assumption is far from reality, it greatly simplifies calculations. This assumption is partially relaxed with the Gamma distribution that models the evolutionary rate heterogeneity among sites. However, other parameters, such as the transition probability matrix, are still assumed to be *i.i.d.* along a sequence. More recently, the CAT model was proposed in which sites along a sequence are divided into K distinct classes, whose evolutionary process is characterized by its own rate matrix (Lartillot and Philippe 2004; Lartillot, Brinkmann, and Philippe 2007; Lartillot and Philippe 2008). CAT is applied in a program called PhyloBayes and has shown its superior resolving power in recent phylogenetic studies (Lartillot and Philippe 2004; Lartillot, Brinkmann, and Philippe 2007; Lartillot and Philippe 2008).

Based on the property of a reversible Markov process, the rate matrix (R) derived from the transition probability matrix (P) can be expressed as the product of two components: the rate parameters Q (also called the exchangeability parameters) and the equilibrium frequencies π (or stationary probabilities). The CAT model assumes that all classes share the same rate parameters Q, but that they have a different set of equilibrium frequencies π for each class. The rate parameters Q can be fixed to the traditional empirical Dayhoff, JTT or WAG matrices to keep computation tractable. The equilibrium frequencies π for each class are estimated from the dataset.

Case studies have shown that the CAT model provides a significantly better fit with data, and that it is more robust against phylogenetic artifacts such as long branch attraction than other models (Lartillot, Brinkmann, and Philippe 2007; Lartillot and Philippe 2008). However, due to its complexity, it needs substantial sequence data to estimate its parameters, and is therefore best suited for use with large (phylogenomic) datasets. Its current implementation uses a Bayesian approach and employs a Markov Chain Monte Carlo (MCMC) technology. The high clade-supporting posterior probabilities provided by MCMC are somehow worrisome because they appear to overestimate the probability that the reconstructed topology represents true evolutionary relationships (Suzuki, Glazko, and Nei 2002; Douady et al. 2003). However, a combination of the Bayesian approach with bootstrapping or jackknifing provides a robust solution, although it is computational expensive - especially for large datasets.

4. Likelihood-based methods for phylogenetic inference

Because phylogenetic inference can be treated as a statistical inference, standard statistical frameworks like least square and likelihood methods can be directly applied. Likelihood methods are most efficient in extracting information compared to least square and others, and the likelihood estimates have a variety of good properties. For example, the estimates convert to the correct value of the parameter (consistency), and have the smallest variance around the true parameter value (efficiency) if the dataset is large enough. Thus, in this study, we focus on likelihood methods for phylogenetic inference. In this section, likelihood function and likelihood-based method for phylogenetic inference are introduced.

4.1 Likelihood function.

The likelihood function (L = Prob(D|M), where D is the data, M is the model) plays a central role in all applications of likelihood-based methods. After an evolutionary model (the tree topology is considered as a parameter of the model) is selected, the likelihood function is used to calculate the probability of a given set of data D for a given tree T (the likelihood value for tree T): L = Prob(D|T, θ), where θ is a vector of parameters for a specified model. Two assumptions are central to computing likelihood values (recent developments have relaxed these assumptions, e.g., (Felsenstein and Churchill 1996)): 1): Evolution at different sites (on the given tree) is independent; 2): Evolution in different lineages is independent. These assumptions significantly simplify the calculation, e.g., based on the first assumption, the likelihood values for a dataset are the product of the L_i

for each sites. Felsenstein developed a practical method for their calculation (Felsenstein 1981).

4.2. Maximum likelihood (ML) method.

The maximum likelihood method aims at identifying the tree with the highest likelihood value. It was first introduced for phylogeny reconstruction by Edwards and Cavalli-Sforza (Edwards and Cavalli-Sforza, 1964) and further developed by Felsenstein (Felsenstein 1981). We can evaluate the likelihood of any given tree T for any given parameter θ; yet the difficulty consists in maximizing the likelihood over all T and all θ. Due to the rapid increase of the number of possible trees with the number of taxa, exhaustive tree search is virtually impossible for real-world datasets. Therefore, a number of heuristic algorithms have been developed, like those implemented in Tree-PUZZLE (Schmidt et al. 2002), PhyML (Guindon and Gascuel 2003b), Treefinder (Jobb, von Haeseler, and Strimmer 2004), RAxML (Stamatakis 2006), etc.

4.3 Bayesian inference (BI) method.

The principle of the BI method is Bayes's theorem:

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}$$

where P(T|D) is the posterior probability of a tree T given data D, P(T) is prior probability of the tree T, the P(D|T) is the likelihood of D given T, the denominator P(D) is the sum of numerators P(D,T) over all possible trees T, and is the quantity that is needed to normalize them so that they add up to 1. The objective of the BI method is to find the tree with a maximum posterior probability that is chosen as the best estimate (Ronquist and Huelsenbeck 2003). However, the denominator (P(D)) in the expression of posterior probability is difficult to compute, as it involves summing over all possible trees. The recent developed Markov chain Monte Carlo (MCMC) methods (Larget and Simon, 1999; Li, Doss and Pearl, 2000) allow to bypass the calculation of P(D) by sampling from the posterior distribution directly.

MCMC is a popular method used for evaluating integrals and solving optimization problems, especially when numerical or other methods cannot be easily applied, such as high dimensional problems. The most common form of MCMC is the Metropolis-Hasting algorithm (Metropolis et al , 1953; Hastings, 1970) (the Gibbs algorithm is a special case of the Metropolis-Hasting, (Geman and Geman, 1984)). The idea is to create a proposal distribution q on the parameter space. Instead of using q to generate a sequence of points sampled from parameter space, q is used to generate a candidate for the next sampled point that will be either accepted or rejected with some probability. If the candidate is rejected, the current point is sampled again. The random acceptance of proposals effectively changes the transition probabilities. Eventually, an appropriate choice of acceptance probabilities will result in a Markov transition matrix q', whose stationary distribution is proportional to the target distribution.

In phylogenetics analyses, MCMC produces a posterior distribution of topologies, and methods are needed to summarize this distribution. Two popular approaches are the maximum posterior probability topology (Rannala and Yang, 1996), and the majority-rule consensus topology (Huelsenbeck et al, 2002). The maximum posterior probability topology is the one with the highest marginal posterior probability. In practice, there might be a large number of different trees in the sampled posterior. The majority-rule consensus topology is the topology with splits that have a marginal posterior probability greater than 0.5.

The BI method shares many fundamental components with ML, like the evolutionary model and the likelihood function. An advantage of the BI method is its capability to handle evolutionary model with high dimensional parameters (like the CAT model) using MCMC method, while ML often fails to do so (Huelsenbeck et al. 2001).

5. Challenges in phylogenetic analysis.

Comparative genomics reveals an enormous heterogeneity among sequences from different species, e.g., different substitution rates, sequence composition and gene content (Lang, Gary and Burger 1999; Burger, Gary and Lang, 2003; Dujon et al, 2004; Xie et al, 2005). This kind of heterogeneity will almost certainly lead to systematic error in phylogenetic analysis; at its extreme, the consequence may be inaccurate (sometimes significantly supported) tree topologies. For instance, in his seminal paper (Felsenstein, 1978), Felsenstein illustrates that when evolutionary rate variation across unrelated lineages is high, they may be incorrectly grouped together in phylogenetic analysis using parsimony method, a phenomenon termed long branch attraction (LBA). It is demonstrated that LBA affects available inference methods without exception, although at varying degrees. With likelihood-based methods, systematic error, like LBA is often due to model violation, because of the model’s unrealistic features (Sullivan and Swofford 2001). In this thesis, LBA is used as a general term to describe systematic error derived from model violations.

Rapid radiation within a short time span represents another challenge (Whitefield and Kjer, 2008). Through incomplete lineage sorting, polymorphisms in an ancestral population can persist through species divergences, resulting in misleading similarities of DNA sequences that do not necessarily reflect population relationships (Pollard et al, 2006). The consequence is the discordance between gene and species trees, and a phylogeny with short unsupported internal branches (Degnan and Rosenberg, 2009). These issues have come into focus because of the growing capacity to generate data sets containing large number of genes for phylogenetic analyses (Delsuc, Brinkmann, and Philippe, 2005).

The following section will describe methods to detect and avoid LBA artifact and rapid radiation.

5.1 The LBA artifact and methods to detect and avoid it

5.1.1 LBA is widespread.

LBA is widespread in phylogenetic inference, at any level of taxonomy. A few examples taken from recent reviews (Philippe 2000; Bergsten 2005) are: the tree of life and the kingdom relationships of Eukaryotes; the class and phylum level of metazoans and plants;

the ordinal level of mammals and birds; and the genus and family level of fish and insects. LBA is also suggested for different data types (DNA, RNA, and amino acid sequences) and data sources alike, including nuclear, mitochondrial and chloroplast datasets.

LBA was first theoretically demonstrated using a four-taxon dataset with the parsimony method. High frequency of parallel change in different species can cause sequence positions to arrive at the same state, undistinguishable from the true phylogenetic signal (Felsenstein 1978b). A number of following studies found that also distance methods suffer seriously from LBA; and that ML methods are least sensitive without eliminating it (Philippe et al. 2000b).

5.1.2 Methods to detect and avoid LBA.

LBA is notoriously difficult to detect and avoid. For example, a fast evolving species may not have a notably long branch in an incorrect phylogeny: when the out-group is distant from other species, fast-evolving species will be attracted to the base of the tree, and their branch length may not be notably long (Philippe 2000). Statistical support values cannot be used as indicator of LBA, as a strong artifact may lead to a well-supported yet incorrect tree topology. Therefore, evidence other than branch length and support values is required to trace LBA (Aguinaldo et al. 1997).

The shape of a tree topology with branch length, in conjunction with the tendency of fast-evolving species to vary position with varying taxon sampling, is a useful indictor for diagnosing LBA. For instance, when a tree is rooted with a distant out-group, fast-evolving lineages are likely attracted to the base (Philippe et al. 2000b). If the relationship among lineages is very different based on different genes, those phylogenies may be due to LBA (to be distinguished from lack of signal and lateral gene transfer). Yet another method to detect LBA is to compare phylogenies and statistical support values with different inference methods. Parsimony and distance methods are more sensitive to LBA than likelihood-based methods. Therefore, if a well-supported grouping with parsimony becomes weakly supported with likelihood methods, this grouping is likely due to LBA.

Another common approach to diagnose LBA uses systematic variation in data sampling (Philippe, Lartillot, and Brinkmann 2005), including either elimination or addition of taxa (i.e., elimination of the fast-evolving, and whenever they become available, addition of slowly evolving species that break up long internal branches), genes (preferably functionally unrelated genes), or sequence positions. The ultimate approach to avoid phylogenetic artifacts is the use of a more realistic (but more complex and computationally more demanding) evolutionary model. Two examples are the Gamma model that takes into account site rate variation, and the CAT model that accounts for account for site-specific features in the evolutionary processes. The application of these two models is known to suppress LBA in given examples (Lartillot, Brinkmann, and Philippe 2007).

5.1.3 LBA caused by compositional heterogeneity.

Most available evolutionary models make the assumption of compositional homogeneity. However, compositional heterogeneity and its effect on phylogenetic analysis has long been recognized and described for both, nucleotide (Hasegawa and Hashimoto 1993) and protein sequences (Foster, Jermiin, and Hickey 1997; Foster and Hickey 1999).

To overcome the artifact caused by compositional heterogeneity at the nucleotide sequence level, the simplest approach consists in RY coding (Phillips, Delsuc, and Penny 2004). A serious drawback of RY coding is loss of phylogenetic information and decrease in phylogenetic resolution. The development of evolutionary models that take nucleotides compositional bias into account are a better alternative, for example, the models developed in (Galtier and Gouy 1998; Foster 2004). For protein coding genes, the common approach consists in analysis at the amino acid level, although proteins sequences are not completely free from compositional bias (Foster, Jermiin, and Hickey 1997; Foster and Hickey 1999).

The new CAT-BP model accounts for variations along lineages by combination of CAT with the non-stationary break point (BP) model (Blanquart and Lartillot 2006; Blanquart and Lartillot 2008). In this combination, equilibrium frequencies change along lineages, with the potential to overcome the effects of compositional bias. It was shows that CAT-BP significantly outperforms the widely used WAG and CAT models in terms of both model fitness and accuracy of phylogenetic inference (Blanquart and Lartillot 2008), yet its application is limited because of its high computational requirements.

5.2 Rapid radiation and incomplete lineage sorting

Although large amounts of data, advanced evolutionary models and newly developed methodologies have become available, many phylogenetic relationships continue to be unresolved. Some examples are relationship among most of metazoan phyla (Rokas, Krüger, and Carroll, 2005), major groups of insects (Whitfield and Kjer, 2008), and the rodent genus (Thomomys) (Belfiore, Liu, and Moritz, 2008). One of the proposed reasons for lacking resolution is rapid radiation.

Detection of rapid radiation is difficult because it is not the only explanation for poorly resolved internal branch. Others include inadequate data, conflict within or among datasets, or loss of phylogenetic signal over time, and inappropriate phylogenetic methods and substitution models (Whitfield and Lockhart, 2007; Rokas and Carroll, 2006). Methods suggested for detecting real ancient radiations include comparison of results from different data types, to detect a potential conflict. The following sections briefly describe these methods.

In the case of rapid radiation, the phylogenetic relationship will not receive significant support using any data type (DNA, RNA, protein, genomic or morphologic data (Mardulyn and Whitfield, 1999)). Applying the realistic and sophisticated evolutionary model will not improve the resolution, compared with the simple model. Thus, comparison of results from different data types and models can help to understand the true reason behind the poor resolution.

The conflicting signals within a dataset may be detected by likelihood mapping (Strimmer and Haeseler, 1997), which analyzes all possible quartets of a dataset, and represents the result in an equilateral triangle. The vertices of the triangle represent three possible tree topologies of a quartet. The bootstrap support values for three trees will determine one point inside the triangle. If there is enough phylogenetic information in the data, then most probability points will fall close to one of the vertices; conversely, datasets containing little phylogenetic information will mainly result in points falling into the center region of the triangle. Likelihood mapping can be obtained using the TreePuzzle package (Schmidt, et al., 2002).

As the time between divergences is shorter due to rapid radiation, different ancestral alleles may be present in the distantly related lineages, this is called incomplete lineage sorting problems (Maddison and Knowles, 2006). Consequently, sequence similarity may not reflect the true evolutionary relationship. In these situations, different genes may suggest different relationships and the underlying species phylogeny will be difficult to resolve (Knowles and Carstens, 2007; Heled and Drummond, 2010).

Incomplete lineage sorting is widespread in closely related speceis phylogenies than deep one, like Fungal phylogeny (Maddison and Knowles, 2006). Distantly related fungal taxa and extinction events decrease the chance that the polymorphisms of ancestral population appear in different lineages, then, reduce the likelihood of incomplete lineage sorting. Thus, incomplete lineage sorting is not a focus of this thesis.

6. Objectives of this study.

Our major objectives are to investigate unresolved phylogenetic issues by taking advantage of new data that were produced in our laboratory, and to develop new methods that increase the accuracy of phylogenetic inference. Phylogenetic questions include the postulated monophyly of Taphrinomycotina, the exact phylogenetic position of Nucleariida, and the relationships among and within major fungal groups, such as zygomycetes and chytridiomycetes.

To elucidate the phylogeny of Taphrinomycotina, we will address the following questions:

- what is the relationship between *Schizosaccharomyces* and other Taphrinomycotina?
- why does the mitochondrial dataset support a different relationship than nuclear dataset?

For the Nucleariida phylogeny, the following questions were asked:

- where is the nucleariid's exact position in the eukaryotic tree?
- is a mitochondrial dataset sufficient to resolve this issue?
- how many sequence positions are required to resolve this question with confidence?

We further attempted to resolve relationships for two less well-known major fungal groups, defined in traditional taxonomies as zygomycetes and chytridiomycetes:

- are they monophyletic or paraphyletic?
- if paraphyletic, can we resolve their branching order with confidence, with the currently available data?

Previous analyses based on single or a small number of genes failed to provide phylogenetic resolution of these questions. In this study, we tackle them by comparing results with large nuclear gene and mitochondrial gene datasets, and by employing likelihood-based methods that are most reliable and robust.

As illustrated by many previous studies and the analysis of Taphrinomycotina using mitochondrial data (this study), phylogenetic artifacts are widespread and difficult to overcome. Recent studies show that heterotachous sites significantly contribute to LBA. We investigate if removal of highly heterotachous (HH) sites from a targeted group of species improves the prediction of correct phylogenetic relationships. For this, we have developed a statistical method that identifies and gradually removes HH sites. The effectiveness of our sequence removal procedure on phylogenetic inference is studied using two published datasets.

References

Adachi, J., and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol **42**:459-468.

Adl, S. M., A. G. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, T. Y. James, S. Karpov, P. Kugrens, J. Krug, C. E. Lane, L. A. Lewis, J. Lodge, D. H. Lynn, D. G. Mann, R. M. McCourt, L. Mendoza, O. Moestrup, S. E. Mozley-Standridge, T. A. Nerad, C. A. Shearer, A. V. Smirnov, F. W. Spiegel, and M. F. Taylor. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol **52**:399-451.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature **387**:489-493.

Baldauf, S. 2008. An overview of the phylogeny and diversity of eukaryotes. JOURNAL OF SYSTEMATICS AND EVOLUTION **46**:263-273.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972-977.

Belfiore, N. M., L. Liu, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus Thomomys (Rodentia: Geomyidae). Syst Biol. 57:294-310.

Bergsten, J. 2005. A review of long-branch attraction. Cladistics **21**:163-193.

Blanquart, S., and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol **25**:842-858.

Blanquart, S., and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol **23**:2058-2071.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology **54**:743-757.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res **31**:1614-1623.

Bullerwell, C. E., J. Leigh, L. Forget, and B. F. Lang. 2003. A comparison of three fission yeast mitochondrial genomes. Nucleic Acids Res **31**:759-768.

Burger, G., M. W., Gray, and B. F. Lang. 2003. Mitochondrial genomes: anything goes. Trends Genet. 19:709-16.Cafaro, M. J. 2005. Eccrinales (Trichomycetes) are not fungi, but a clade of protists at the early divergence of animals and fungi. Mol Phylogenet Evol **35**:21-34.

Carr, M., B. S. Leadbeater, R. Hassan, M. Nelson, and S. L. Baldauf. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. Proc Natl Acad Sci U S A **105**:16641-16646.

Cavalier-Smith, T. 2004. Only six kingdoms of life. Proc Biol Sci **271**:1251-1262.

Cavalier-Smith, T., and E. E. Chao. 2003. Phylogeny and classification of phylum Cercozoa (Protozoa). Protist **154**:341-358.

Dayhoff, M., RM Schwartz and BC Orcutt. 1978. A model for evolutionary change in proteins. Pp. 345-352 *in* M. Dayhoff, ed. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, New York.

Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332-40.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361-75.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol **20**:248-254.

Dujon, B., D. Sherman D, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E.

Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker,J. L. Souciet. 2004. Genome evolution in yeasts. Nature. 430:35-44.

Edman, J., J. Kovacs, H. Masur, D. Santi, H. Elwood, and M. Sogin. 1988. Ribosomal RNA sequence shows Pneumocystis carinii to be a member of the fungi. Nature **334**:519-522.

Edwards, A. W. F. and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. pp. 67-76 in Phenetic and Phylogenetic Classsification, ed. V. H. Heywood and J. McNeill. Systematics Association Publication No. 6, London.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

Felsenstein, J., and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol **13**:93-104.

Fitzpatrick, D. A., M. E. Logue, J. E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol **6**:99.

Forget, L., J. Ustinova, Z. Wang, V. A. Huss, and B. F. Lang. 2002. Hyaloraphidium curvatum: a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi. Mol Biol Evol **19**:310-319.

Foster, P. G. 2004. Modeling compositional heterogeneity. Systematic Biology **53**:485-495.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol **48**:284-290.

Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol **44**:282-288.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol **18**:866-873.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol **15**:871-879.

Geman, S. and D. Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Patt Ana Mach Intell 6: 721–741.

Gonnet, G. H., M. A. Cohen, and S. A. Benner. 1992. Exhaustive matching of the entire protein sequence database. Science **256**:1443-1445.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52**:696-704.

Gunderson, J. H., H. Elwood, A. Ingold, K. Kindle, and M. L. Sogin. 1987. Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. Proc Natl Acad Sci U S A **84**:5823-5827.

Hastings, W.K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57: 97–109.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

Hawksworth, D. L. 2001. The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycol. Res. **105**:1422-1432.

Heled, J. and A. J. Drummond 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570-80.

Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res **111**:509-547.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310-2314.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F.

Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature **443**:818-822.

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

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol **4**:18.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.

Keeling, P. 2009. Five questions about microsporidia. PLoS Pathog **5**:e1000489.

Keeling, P. J. 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet Biol **38**:298-309.

Keeling, P. J., M. A. Luker, and J. D. Palmer. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol Biol Evol **17**:23-31.

Kim, E., and L. E. Graham. 2008. EEF2 analysis challenges the monophyly of Archaeplastida and Chromalveolata. PLoS ONE **3**:e2621.

King, N., M. J. Westbrook, S. L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K. J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J. B. Lyons, A. Morris, S. Nichols, D. J. Richter, A. Salamov, J. G. Sequencing, P. Bork, W. A. Lim, G. Manning, W. T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I. V. Grigoriev, and D. Rokhsar. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature **451**:783-788.

Knowles, L. L. and B. C. Carstens. 2007. Delimiting species without monophyletic gene trees. Syst Biol. 56:887-95.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr Biol **12**:1773-1778.

Lang, B. F., M. W. Gray, and G. Burger. 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet. 33:351-97.

Lara, E., D. Moreira, and P. Lopez-Garcia. 2009. The Environmental Clade LKM11 and Rozella Form the Deepest Branching Clade of Fungi. Protist.

Larget, B. and D. L. Simon, 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16: 750-759.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol **7 Suppl 1**:S4.

Lartillot, N., and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci **363**:1463-1472.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095-1109.

Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol Biol Evol **25**:1307-1320.

Le, S. Q., N. Lartillot, and O. Gascuel. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci **363**:3965-3976.

Leigh, J., E. Seif, N. Rodriguez, Y. Jacob, and B. F. Lang. 2003. Fungal evolution meets fungal genomics. Pp. 145-161 *in* D. K. Arora, ed. Handbook of Fungal Biotechnology. Marcel Dekker Inc., New York.

Li, S. Y. Pearl, D. K. and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J Am Stat Ass 95:493-508.

Li, W.-H. 1997. Molecular evolution. Sinauer, Sunderland MA.

Lio, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res **8**:1233-1244.

Liu, Y. J., M. C. Hodson, and B. D. Hall. 2006. Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. BMC Evol Biol **6**:74.

Maddison, W.P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 2006 Feb;55(1):21-30.

Mardulyn, P and J. B. Whitfield. 1999. Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera; Braconidae): evidence of a high diversification rate in this group of parasitoids. Mol Phylogenet Evol. 12:282-94.

Marshall, W. L., G. Celio, D. J. McLaughlin, and M. L. Berbee. 2008. Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from marine invertebrate digestive tracts. Protist **159**:415-433.

McLaughlin, D. J., D. S. Hibbett, F. Lutzoni, J. W. Spatafora, and R. Vilgalys. 2009. The search for the fungal tree of life. Trends Microbiol **17**:488-497.

Medina, M., A. G. Collins, J. W. Taylor, J. W. Valentine, J. H. Lipps, L. Amaral-Zettler, and M. L. Sogin. 2003. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. International Journal of Astrobiology **2**:203-211.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. J Chem Phys 21: 1087–1092.

Moreira, D., H. Le Guyader, and H. Philippe. 2000. The origin of red algae and the evolution of chloroplasts. Nature **405**:69-72.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. Pp. 1-27 *in* S. S. G. a. J. Yackel, ed. Statistical decision theory and related topics. Academic Press, New York.

Nishida, H., and J. Sugiyama. 1993. Phylogenetic relationships among *Taphrina, Saitoella*, and other higher fungi. Mol Biol Evol **10**:431-436.

Paquin, B., M. J. Laforest, L. Forget, I. Roewer, Z. Wang, J. Longcore, and B. F. Lang. 1997. The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. Curr Genet **31**:380-395.

Parfrey, L. W., E. Barbero, E. Lasser, M. Dunthorn, D. Bhattacharya, D. J. Patterson, and L. A. Katz. 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS Genet **2**:e220.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol **53**:711-723.

Philippe, H. 2000. Opinion: long branch attraction and protist phylogeny. Protist **151**:307-316.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol **22**:1246-1253.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc Biol Sci **267**:1213-1221.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol **5**:50.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol **21**:1455-1458.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting, PLoS Genet. 2(10):e173.

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43: 304-11.

Rodriguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe, and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol **15**:1325-1330.

Rodriguez-Ezpeleta, N., H. Brinkmann, G. Burger, A. J. Roger, M. W. Gray, H. Philippe, and B. F. Lang. 2007a. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol **17**:1420-1425.

Rodriguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe. 2007b. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology **56**:389-399.

Rokas, A and S. B. Carroll. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Rokas, A., D. Krüger, and S. B. Carroll. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933-8.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Ruiz-Trillo, I., C. E. Lane, J. M. Archibald, and A. J. Roger. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts Capsaspora owczarzaki and Sphaeroforma arctica. J Eukaryot Microbiol **53**:379-384.

Ruiz-Trillo, I., A. J. Roger, G. Burger, M. W. Gray, and B. F. Lang. 2008. A phylogenomic investigation into the origin of metazoa. Mol Biol Evol **25**:664-672.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504.

Schwarzott, D., C. Walker, and A. Schussler. 2001. Glomus, the largest genus of the arbuscular mycorrhizal fungi (Glomales), is nonmonophyletic. Mol Phylogenet Evol **21**:190-197.

Seif, E., J. Leigh, Y. Liu, I. Roewer, L. Forget, and B. F. Lang. 2005. Comparative mitochondrial genomics in zygomycetes: bacteria-like RNase P RNAs, mobile elements and a close source of the group I intron invasion in angiosperms. Nucleic Acids Res **33**:734-744.

Shalchian-Tabrizi, K., M. A. Minge, M. Espelund, R. Orr, T. Ruden, K. S. Jakobsen, and T. Cavalier-Smith. 2008. Multigene phylogeny of choanozoa and the origin of animals. PLoS ONE **3**:e2098.

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, S. Stefanovic, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci **9**:477-483.

Spatafora, J. W., G. H. Sung, D. Johnson, C. Hesse, B. O'Rourke, M. Serdani, R. Spotts, F. Lutzoni, V. Hofstetter, J. Miadlikowska, V. Reeb, C. Gueidan, E. Fraker, T. Lumbsch, R. Lucking, I. Schmitt, K. Hosaka, A. Aptroot, C. Roux, A. N. Miller, D. M. Geiser, J. Hafellner, G. Hestmark, A. E. Arnold, B. Budel, A. Rauhut, D. Hewitt, W. A. Untereiner, M. S. Cole, C. Scheidegger, M. Schultz, H. Sipman, and C. L. Schoch. 2006. A five-gene phylogeny of Pezizomycotina. Mycologia **98**:1018-1028.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Steenkamp, E. T., J. Wright, and S. L. Baldauf. 2006. The protistan origins of animals and fungi. Mol Biol Evol **23**:93-106.

Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol **4**:35.

Strimmer, K and A. von Haeseler. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. PNAS. 94:6815-9.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology **50**:723-729.

Sumathi, J. C., S. Raghukumar, D. P. Kasbekar, and C. Raghukumar. 2006. Molecular evidence of fungal signatures in the marine protist Corallochytrium limacisporum and its implications in the evolution of animals and fungi. Protist **157**:363-376.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A **99**:16138-16143.

Tanabe, Y., M. M. Watanabe, and J. Sugiyama. 2005. Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): Insights from molecular phylogenetics. J Gen Appl Microbiol **51**:267-276.

Taylor, J., J. Spatafora, K. O'Donnell, F. Lutzoni, T. James, D. Hibbett, D. Geiser, T. Bruns, and M. Blackwell. 2004. The Fungi. Pp. 171-194 *in* M. J. D. Joel Cracraft, ed. Assembling the Tree of Life. Oxford University Press, New York.

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

Thomarat, F., C. P. Vivares, and M. Gouy. 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.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18**:691-699.

Whitfield, J. B. and K. M. Kjer. 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. Annu Rev Entomol. 53:449-72.

Whitfield, J.B. and P. J. Lockhart. 2007. Deciphering ancient rapid radiations. Trends Ecol Evol. 22:258-65.

Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 434:338-45.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution:367-372.

Yoon, H. S., J. Grant, Y. I. Tekle, M. Wu, B. C. Chaon, J. C. Cole, J. M. Logsdon, Jr., D. J. Patterson, D. Bhattacharya, and L. A. Katz. 2008. Broadly sampled multigene trees of eukaryotes. BMC Evol Biol **8**:14.

Zettler, L. A. A., T. A. Nerad, C. J. O'Kelly, and M. L. Sogin. 2001. The nucleariid amoebae: more protists at the animal-fungal boundary. J Eukaryot Microbiol **48**:293-297.

Chapter 2 Comparative mitochondrial genomics in zygomycetes: bacteria-like RNase P RNAs, mobile elements and a close source of the group I intron invasion in angiosperms

Elias Seif[1], Jessica Leigh[2], Yu Liu[1], Ingeborg Roewer[3], Lise Forget[1] and B. Franz Lang[1,*]

[1]*Program in Evolutionary Biology, Canadian Institute for Advanced Research; Centre Robert Cedergren, Département de Biochimie, Université de Montréal 2900 Boulevard Edouard-Montpetit, Montréal, Québec, Canada H3T 1J4*

[2] *Department of Biochemistry and Molecular Biology, Dalhousie University Halifax (Nova Scotia), Canada B3H 4H7*

[3] *National Research Council of Canada, Plant, Biotechnology Institute 110 Gymnasium Place, Saskatoon, SK, Canada S7N 0W9*

[*]To whom correspondence should be addressed. Tel: +1 514 343 5842; Fax: +1 514 343 2210;

ABSTRACT

To generate data for comparative analyses of zygomycete mitochondrial gene expression, we sequenced mtDNAs of three distantly related zygomycetes, *Rhizopus oryzae*, *Mortierella verticillata* and *Smittium culisetae*. They all contain the standard fungal mitochondrial gene set, plus *rnpB*, the gene encoding the RNA subunit of the mitochondrial RNase P (mtP-RNA) and *rps3*, encoding ribosomal protein S3 (the latter lacking in *R. oryzae*). The mtP-RNAs of *R. oryzae* and of additional zygomycete relatives have the most eubacteria-like RNA structures among fungi. Precise mapping of the 5' and 3' termini of the *R. oryzae* and *M. verticillata* mtP-RNAs confirms their expression and processing at the exact sites predicted by secondary structure modeling. The 3' RNA processing of zygomycete mitochondrial mRNAs, SSU-rRNA and mtP-RNA occurs at the C-rich sequence motifs similar to those identified in fission yeast and basidiomycete mtDNAs. The C-rich motifs are included in the mature transcripts, and are likely generated by exonucleolytic trimming of RNA 3' termini. Zygomycete mtDNAs feature a variety of insertion elements: (i) mtDNAs of *R. oryzae* and *M. verticillata* were subject to invasions by double hairpin elements; (ii) genes of all three species contain numerous mobile group I introns, including one that is closest to an intron that invaded angiosperm mtDNAs; and (iii) at least one additional case of a mobile element, characterized by a homing endonuclease insertion between partially duplicated genes [Paquin,B., Laforest,M.J., Forget,L., Roewer,I., Wang,Z., Longcore,J. and Lang,B.F. (1997) Curr. Genet., 31, 380–

395]. The combined mtDNA-encoded proteins contain insufficient phylogenetic signal to demonstrate monophyly of zygomycetes.

INTRODUCTION

Fungi constitute a huge group of highly diverse organisms, including some of the most-studied and best-understood eukaryotic model systems: ‘baker's yeast’ (*Saccharomyces cerevisiae*), fission yeast (*Schizosaccharomyces pombe*), and the filamentous euascomycetes *Neurospora crassa* and *Aspergillus nidulans*. These species all belong to the Ascomycota. Substantially fewer scientific studies have been performed in members of the sister phylum Basidiomycota and very little is known in the remaining two phyla, Zygomycota and Chytridiomycota, often classified as ‘lower fungi’. This expression is a taxonomically vague concept borrowed from Aristotle's philosophy, visioning directed evolution from the simple (primitive, low) to the highly complex. The misnomer is most evident in the ‘higher’ ascomycetes and basidiomycetes, which evolve toward microscopic, unicellular and genetically simplified yeast-like organisms in some lineages, and toward morphologically complex, gene-rich and biochemically versatile multicellular organisms in others.

Although the availability of complete nuclear and mitochondrial sequences of more than a dozen ascomycetes provides a strong basis for biochemical investigations, only a few complete mitochondrial sequences are known from chytrids, and none from zygomycetes, a situation that has motivated the work presented here. In fact, the number of zygomycete nuclear gene sequences (mostly rRNA sequences) is so limited that it is impossible to determine with confidence whether or not Zygomycota is a monophyletic taxon (1-3). The lack of resolution in these analyses is consistent with estimates that even the combined LSU- and SSU-rRNA would contain far too little information to resolve many fungal phylogenetic relationships with confidence (4). This dataset is at most sufficient for resolving fungal inferences below the phylum level (5,6).

Sequencing complete mtDNAs from several zygomycetes might be a first step in remediating this situation. Mitochondrial phylogenies can be based on up to 13 protein sequences, and have been shown to resolve deep divergences in the fungal and animal lineages (7-9). For instance, Alexopolous *et al.* (10)indicate that ‘additional study is needed to determine whether the class (Trichomycetes) is a monophyletic group belonging to Zygomycota, or merely a collection of orders grouped together on the basis of a unique shared habitat’. Molecular phylogenies based on rRNA sequences were successful in moving the order Amoebidiales away from zygomycetes, although they were then placed with mesomycetozoan rotists, whose phylogenetic affiliation was unresolved and controversial (11). Only subsequent analysis with multiple mitochondrial proteins placed this group as the closest relative of animal, with high confidence (9). This example clearly illustrates the requirement of multi-gene datasets with at least several thousand amino acid positions, for resolution of trees at the kingdom level.

Furthermore, zygomycete mtDNAs are of considerable interest for comparative gene expression studies: our preliminary data indicated the presence of a mitochondrial *rnpB* gene, which encodes the RNA subunit of RNase P, the enzymatically active part of an endonuclease (ribonucleo-protein) responsible for tRNA maturation. In mitochondria, the size and sequence of the RNA subunit varies substantially, which has considerably complicated its identification. The gene is apparently absent from all completely sequenced basidiomycete and chytridiomycete mtDNAs, and presents in some ascomycetes (12); to date, there are no published data on zygomycete mitochondrial RNase P (mtP-RNA).

To help remediate the lack of data for phylogenetic inferences, and to facilitate biochemical investigations and comparative mitochondrial genome analyses, we have sequenced mtDNAs from three distantly related zygomycetes, the Mucorales *Rhizopus oryzae*, the Mortierellales *Mortierella verticillata* and the Harpellales *Smittium culisetae*. In this article, we compare their mitochondrial genomes (gene content, gene organization, genetic code and widely conserved 3' RNA processing sites). We will then present secondary structure models and expression data for seven newly identified zygomycete mtP-RNAs. Finally, we will test whether zygomycetes are monophyletic, and provide evidence that the group I introns invasion of *cox1* gene in angiosperms originated in a zygomycete close to Rhizopus.

MATERIALS AND METHODS

Strains and culture conditions

The various zygomycete strains were obtained from Kerry O'Donnell (National Center for Agricultural Utilization Research, Peoria, IL; NRRL), R.W. Lichtwardt (Department of Botany, University of Kansas, Laurence, KS; RWL) and Carolyn Babcock (Canadian Collection of Fungal Cultures, Ottawa; DAOM). All strains, *R. oryzae* (DAOM 148428, previously designated as *Rhizopus stolonifer*), *R.stolonifer* (DAOM 194667), *Rhizopus oligosporus* (NRRL 2710), *M.verticillata* (NRRL 6337), *Radiomyces spectabilis* (NRRL 2753), *Mucor mucedo* (NRRL 3635) and *S. culisetae* (strain 18-3; R.W. Lichtwardt), were grown in YG medium consisting of 0.5% yeast extract and 3% glycerol. Liquid cultures of 500 ml in 2 L Erlenmeyer flasks were grown at room temperature under gentle shaking (~ 100 r.p.m.).

DNA and RNA extractions

For mtDNA and RNA extractions, the cells were broken mechanically, and a mitochondrial fraction was isolated by differential centrifugation (8). This fraction was lysed in the presence of 1% SDS and 100 µg/ml proteinase K at 50°C for 1 h, and after phenol–chloroform extraction, the nucleic acids were precipitated with ethanol. For RNA purifications, the high molecular weight RNA fraction was precipitated with 2 M LiCl, redissolved in RNase-free water and ethanol-precipitated. MtDNAs from all zygomycete strains were purified from total cellular nucleic acids by Cesium chloride/bisbenzimide density gradient centrifugation.

Cloning and sequencing of complete mtDNAs

Library construction and DNA sequencing followed previously published protocols (8). Briefly, mtDNAs were physically sheared by nebulization (13), and a size fraction of 1300–4000 bp was recovered after agarose gel electrophoresis. The DNA was incubated with a mixture of T7 DNA polymerase and *Escherichia coli* DNA polymerase I (the Klenow fragment) to generate blunt ends, and then cloned into the EcoRV cloning site of the phagemid pBFL6 (B. F. Lang, unpublished data). Recombinant plasmids containing mtDNA inserts were identified by colony hybridization using mtDNA as a probe. Clones contained in the random libraries were sequenced to ~ 8-fold coverage, and remaining gaps were closed by primer walking or sequencing of PCR-amplified DNA fragments. The expected quality of the sequenced mtDNAs is <1 error in 10 000 bp.

The mtDNA sequences of *R. oryzae*, *M. verticillata* and *S. culisetae* have been deposited in GenBank (accession nos AY863212, AY863211 and AY8632133, respectively).

PCR amplification of *rnpB* genes

Mitochondrial *rnpB* genes of *R. stolonifer*, *M. mucedo*, *R. spectabilis* and *R. oligosporus* were PCR-amplified from ~100 ng of the respective mtDNAs in a 50 μl reaction mixture [200 μM dNTP, 2.5 mM $MgCl_2$, 2 nM of primers, 5 μl of 10x buffer and 3 U of DNA polymerases mixture from the Expand high fidelity kit (Roche Catalog no. 1732650) and degenerate primers]. The annealing temperature of the PCR amplification was 50°C. Sequences of the degenerate primers are 5'-GTAATGGCAGCATACTAGACTCAT-3' and

5'-TTGAACTCCCAAGTTTTATGTATG-3'. The amplification products were cloned and sequenced. The *rnpB* sequences of *R. stolonifer*, *R. oligosporus*, *M. mucedo* and *R. spectabilis* have been deposited in GenBank (accession nos AY861429, AY861440, AY861441, and AY861442, respectively).

RNA circularization by ligation and RT–PCR

RNA ligation of mtP-RNAs, followed by RT–PCR amplification, was performed to determine the precise 5' and 3' termini of RNase P RNAs according to the previously published protocols (12). The primers are 5'-CTCTTATAGGATAATACAAAGTTG-3' and 5'-GGCCGAAGAATAAAGAGGGA-3' for *M.verticillata*, and 5'-ACCCTAATTTTCATT-AGATATTT-3' and 5'-AATCCTTAGTAAGGATAGCTT-3' for *R.oryzae*.

RT–PCR of mtP-RNA from *R.oryzae*

Mitochondrial RNA of *R. oryzae* was treated with DNase I and extracted with phenol/chloroform, until no genomic DNA could be amplified by PCR using the mtP-RNA-specific primers 5'-TTCTTAGAGTTAAATAAGCC-3' and 5'-TTGGAGGAAAGTCCG-GG-3'. Following these treatments, we amplified the mtP-RNA by RT–PCR as described above using a sample without reverse transcription as negative control. Following amplification and separation on a 0.8% agarose gel, the resulting DNA fragment was cloned and sequenced.

S1 mapping

DNA oligonucleotides and a 10 bp DNA ladder (Invitrogen 10821-015) were labeled at their 3' termini with ddATP-^{32}P (Amersham PB10235) and terminal deoxynucleotidyl transferase (MBI Fermentas EP0161) according to the manufacturer's recommendations.

A total of 100 000 c.p.m. of gel-purified labeled oligonucleotides were hybridized to ∼10 μg of total RNA, as described in the protocol by Hahn and Breeden (http://www.fhcrc.org/labs/hahn/methods/mol_bio_meth/s1_oligo_probe.html), and S1 nuclease digestions were carried out at 37°C for 30 min, after the addition of 20 U of S1 nuclease and the buffer provided by the manufacturer (MBI Fermentas EN0321). The product was then ethanol-precipitated and dissolved in 4 μl of RNase-free water. An aliquot of 2 μl of the product was mixed with 2 μl of the dye solution provided with the 10 bp ladder (Invitrogen 10821-015), denatured at 75°C and loaded on a 7% polyacrylamide denaturing sequencing gel. The following oligonucleotides were used:

SSU-rRNA, 5'-AAATAAAGGGTTTAATATATTGGGAGGGACTTATTGTCCC
CCCGGTAATAACCATTCAGCCACTCGTTCCCGAACGGCT-3'
cox1 mRNA, 5'-CAGATTCTAAGGGGGTGTTATATTATTAATTTAATTAAGA
TTGAACTGGTAATGAATTTACAGTATGGA-3'.

Phylogenetic inference

Mitochondrial protein sequences from all completely sequenced zygomycetes, chytridiomycetes and basidiomycetes were included in phylogenetic inference (for species names and GenBank accession nos, see legend of Figure 5). Protein sequences from apocytochrome *b* (Cob), as well as 7 subunits of NADH dehydrogenase (Nad), 3 cytochrome *c* oxidase (Cox) subunits and 2 ATP synthetase (Atp) subunits were aligned with Muscle (14), concatenated and trimmed with Gblocks [using default parameters (15)] to remove ambiguously aligned regions. The resulting alignment contained 2890 aligned positions. Maximum likelihood (ML) phylogenies were inferred from this alignment using both PHYML (16) and IQPNNI (17); ML bootstrap support was determined based on 100 replicates, using both programs. All phylogenies were inferred using the JTT amino acid substitution model and gamma distribution correction for variation of rates across sites.

Phylogenetic inference based on concatenated datasets can lead to tree reconstruction artefacts, such as long-branch attraction, resulting from differences in relative evolutionary rates between genes [sometimes referred to as the covarion-like behavior of different genes (18)]. Ideally, during the ML tree search procedure, nuisance parameters, such as branch lengths and the Γ distribution shape parameter (α), could be optimized separately for each gene; however, this method results in the estimation of many more parameters, potentially more than can be statistically justified. For this reason, we partitioned the dataset into four functional categories: Atp (253 positions), Cob (361 positions), Cox (876 positions) and Nad (1400 positions). The additional parameters included in this partitioned dataset were justified using the χ^2 test ($P < 0.0001$). Using this partitioned dataset, a topology was inferred from this alignment using MrBayes, with branch lengths and α parameter unlinked across partitions (19). In addition, a ML tree was determined from this partitioned dataset using an adaptation of the method of (18). All possible tree topologies were generated, with constraints of groups that received at least 95% bootstrap support under ML, using both IQPNNI and PHYML (Supplementary Figure 1). Log-likelihoods were calculated separately for each partition using PHYML, under each of the topologies, and the sum over all partitions was calculated for each tree. The tree found to maximize the sum log-likelihood of the dataset was taken to be the ML tree (this method is referred to henceforth as separate analysis). Bootstrap support was also calculated using both of these methods, based on 100 replicates. Likelihood ratio tests were performed using genewise optimized site likelihoods, given the 99 tree topologies described in Supplementary Figure 1, and using Tree-Puzzle to generate sitewise likelihoods (20) along with CONSEL (21).

Additionally, phylogenetic analysis of closely related fungal, green algal and plant intronic open reading frames (ORFs) of the same *cox1* intron (for species names and GenBank accession nos, see legend of Figure 5) were carried out. Sequences were aligned using Muscle and trimmed with Gblocks, and the resulting alignment was manually refined. A ML phylogeny was inferred using IQPNNI, and ML bootstrap support was determined based on 100 replicates.

RESULTS AND DISCUSSION

Genes in addition to the standard fungal set, *rnpB* and *rps3*

The mitochondrial genomes of the zygomycetes *R. oryzae* (previously listed incorrectly as *R. stolonifer*), *M. verticillata* and *S. culisetae* were completely sequenced. Like most other fungal mtDNAs, they map as circular molecules (Figure 1), although they are most likely organized as linear multimeric concatamers *in vivo*, as in other fungi (22). mtDNAs of the three species carry the basic fungal set of genes (Table 1), and encode a full set of tRNAs [only *trnI*(cau) is absent in *R. oryzae*], the RNA component of mitochondrial RNase P (*rnpB*) and a ribosomal protein (*rps3*; lacking in *R. oryzae*). MtDNA-encoded *rps3* has previously been identified in several ascomycetes and in one chytridiomycete (*Allomyces macrogynus*; Blastocladiales), but not in other chytridiomycete orders, including Monoblepharidales, Spizellomycetales and Chytridiales. It has been proposed that this gene has been lost independently three times from opisthokont mitochondrial genomes: in the chytridiomycete lineage, in the animal lineage (23) and now also in one of the three zygomycetes presented here. We assume that, in all cases, *rps3* has been transferred to the nuclear genome, like other ribosomal protein genes missing in fungal mitochondrial genomes (24).

The sizes of zygomycete mtDNAs are within a close range of 54–58 kb (Figure 1). Genes are encoded on both strands, but are not as tightly packed as in animals and some ascomycetes: only 40.6% of mtDNAs in *R. oryzae*, 43.1% in *M. verticillata* and 35.3% in *S. culisetae* are coding. Nonetheless, the coding regions of *nad2/nad3* and *nad4L/nad5* of *R. oryzae*, respectively, overlap by 1 nt (i.e. the last nucleotide position of the UAA stop codon of the upstream gene is the first nucleotide of the AUG start codon of the downstream gene; Figure 1). No conservation of mitochondrial gene order is observed between these species.

First steps toward a derived genetic code in zygomycetes

Both *R. oryzae* and the fast-evolving *S. culisetae* have retained the standard translation code for protein coding genes, a trait inherited from the eubacterial ancestors of mitochondria. However, *M. verticillata* reassigns two UGA 'stop' codons as tryptophan, once each in *nad3* and *nad4*. UGA(Trp) codons are also present in the *S. culisetae* intronic ORF283 and ORF248, both encoding group I introns homing endonucleases of the LAGLI-DADG type. UGA(Trp) at amino acid position 237 of ORF248 is part of a distinctive, highly conserved sequence motif of this class of endonuclease, strongly suggesting its translation as tryptophan. It is possible that the presence of this UGA(Trp) is a vestige of horizontal intron transfer from a fungus adapted to this translation code. In fact, according to our phylogenetic analyses of intron endonucleases (Figure 5B), *S. culisetae* ORF248 is closely related to ORF313 of *Podospora anserina*, which makes preferential use of UGA(Trp) in its genes and intronic ORFs. Like in the fission yeast *S. pombe* and the basidiomycete *Schizophyllum commune*, the mtDNAs of *S. culisetae* and *M. verticillata* do not encode *trnW*(uca), which would efficiently recognize both UGA and UGG tryptophan codons. We assume that in all these cases, UGA codons are (albeit inefficiently) decoded by *trnW*(cca) (25, 26). However, it cannot be excluded that, alternatively, the C in the wobble position of the anticodon is either modified or partially edited to allow efficient recognition of UGA(Trp) codons.

The zygomycete mtDNAs described here encode complete sets of tRNAs sufficient to recognize all encountered codons (for codon usage, see Supplementary Table 1S). *R. oryzae* does not have *trnI*(cau); however, ATA(Ile) codons are absent in standard mitochondrial genes, although they occur in intronic ORFs. Incidentally, a strikingly similar scenario exists in *Schizosaccharomyces octosporus*. It has been suggested (27) that either (i) the tRNA required for translation of ATA(Ile) is imported from the cytoplasm to recognize these codons or (ii) the intronic ORFs are neither translated nor required for intron splicing. A further explanation is that these codon positions are recognized by other tRNAs at low efficiency, resulting in amino acid misincorporation, which might be permissible at poorly conserved amino acid positions of proteins.

Whatever the mechanism of codon recognition, we suggest that such unexpected codon usage in intronic ORFs reflects horizontal intron transfer from species that are adapted to the use of UGA(Trp) and/or ATA(Ile). This codon usage is common in fungi and several other eukaryotic lineages.

Eubacteria-like mtP-RNAs in zygomycete mitochondria

Mitochondrial *rnpB* genes (encoding the mitochondrial RNA subunit of RNase P, mtP-RNA) were identified by *in silico* analysis in all three zygomycetes using the previously described procedures (12), and their RNA secondary structures were modeled by phylogenetic comparative analysis (Figure 2) (see also http://megasun.bch.umontreal.ca/People/lang/rnpB/). The presence of this gene in all three zygomycetes is striking, because outside fungi, *rnpB* is only present in the green alga *Nephroselmis olivacea* (28) and in various jakobids (B. F. Lang and E. Seif, unpublished data), including *Reclinomonas americana* (29). Within fungi, it is only present in some ascomycetes, but absent in basidiomycetes and chytridiomycetes (12).

The inferred size of the *S. culisetae* mtP-RNA is 145 nt, close to the shortest known example (140 nt, in *Saccharomycopsis fibuligera*) (30). Most remarkably, the highly reduced mtP-RNA structures of *S. culisetae* and budding yeasts are almost identical (Figure 2), perfectly matching the minimum consensus secondary structure of fungal mtP-RNAs (12). In contrast, *rnpB* from *M. verticillata* and *R. oryzae* are the largest genes of this class ever identified (980 and 830 bp, respectively), even larger than *rnpB* genes studied in bacteria (31). In addition, the zygomycete mitochondrial RNA secondary structures are the most bacteria-like among fungi.

To verify the expression of the *R. oryzae* and *M. verticillata* genes, we determined their precise 5' and 3' ends by sequencing RT–PCR products of circularized mtP-RNAs (Figure 2). The 3' end of *M. verticillata* is 9 nt longer than anticipated, elongated by a cytidine-rich stretch of sequence. A similar extension is located at the 3' terminus of *Schizosaccharomyces octosporus* mtP-RNA (12) and downstream of protein coding genes in a variety of fungi (see below; Figure 3). The 5' end of *M. verticillata*, and both the 5' and 3' termini of *R. oryzae* mtP-RNAs, match the proposed secondary structure model and reveal little heterogeneity of mtP-RNA termini (Figure 2).

In evolutionary terms, zygomycete mtP-RNA structures cover an unprecedented wide range of intermediate stages in loss of RNA structural elements. The mtP-RNAs from *R. oryzae* and closely related species have the most bacteria-like secondary structures, containing almost all structural elements of the bacterial minimum secondary structure consensus (32). They are followed by *M. verticillata*, whose structure closely resembles the more derived mtP-RNA of the ascomycete *Taphrina deformans* (12). Finally, the tiny, yeast-like mtP-RNA molecule of *S. culisetae* has no P2 helix, which is otherwise omnipresent in mtP-RNAs. Note that this helix is also absent in *M. verticillata*, potentially indicating its loss in a common ancestor.

The most bacteria-like fungal mtP-RNA secondary structure is that of *R. oryzae*, only lacking P13, P14 and P19, which are otherwise only present in the protist mtP-RNAs of *N. olivacea* and *R. americana*. The large size of the *R. oryzae* and *M. verticillata* mtP-RNAs is due to insertions at the J5-15 and J5-18 junctions, respectively, and in the P12 helix. In order to determine whether these regions are conserved structural elements or more variable insertion elements or introns, we amplified the cDNA sequence of the *R. oryzae* mtP-RNA, and the genomic sequences from the closely related Mucorales *M. mucedo*, *R. spectabilis*, *R. oligosporus* and *R. stolonifer*. Figure 2 shows that the insertion sequences can be folded into double hairpin structures [DHEs (33)]. Because the cDNA sequence of the *R. oryzae* mtP-RNA is identical to the genomic sequence, these variable regions are not introns. Furthermore, the insertion points and sizes of these regions vary substantially, indicating that they have been acquired recently and independently. Their presence in mtP-RNAs pinpoints structural regions that are likely not critical for RNase P activity. An analogous situation has been described in some cyanobacteria, where P-RNAs contain short tandem repeats that increase the length of helix P12. Site-directed mutagenesis experiments have shown that this helix is not required for catalytic activity *in vitro*, implying that it is also unlikely to be crucial for the *in vivo* activity (34).

Conserved C-rich motifs in mRNAs and SSU-rRNA

Small C-rich clusters are present downstream of mitochondrial protein- and SSU-rRNA coding regions, in all three species (Figure 3). The consensus sequence of this motif varies only slightly among zygomycetes. It also exists in basidiomycetes and in fission yeasts, pointing to a shared function. Mapping of the 3' end of the *M. verticillata* mtP-RNA (which also terminates with this motif as discussed above; Figure 2), of *cox1* and the SSU-rRNA of *R. oryzae* shows that these C-rich motifs are the site of 3' RNA processing, and are retained in the mature RNA molecules. The presence of ragged 3' ends (Figure 3) indicates that these are generated by an exonuclease trimming mechanism. Similar observations have been made in fission yeasts (12, 27). This mechanism resembles that of nuclear and viral RNAs terminating in polyuridine motifs that serve as a binding site for the La protein implicated in RNA protection against exonucleases [reviewed in (35); homologs of La are known from a range of fungi and animals].

More instances of mobile endonuclease elements?

As reported earlier, the chytridiomycete fungus *A. macrogynus* has a novel mtDNA insertion element whose sequence is absent in the close relative, *Allomyces arbusculus*. This element consists of a duplicate C-terminus of a foreign *atp6* gene, plus an ORF encoding an endonuclease that is responsible for its mobility (36). The inserted *atp6* portion is fused in phase with the resident gene (Figure 4), reconstituting an obviously functional hybrid gene of standard length and amino acid conservation. In fact, it has been shown that homing endonucleases can be mobile even independent of introns and genes (37), and that they are capable of carrying genetic material from one site to another when they migrate.

Intriguingly, similar gene hybrids are present in the mtDNAs of *R. oryzae* (*atp9*-C*) and *M. verticillata* (*cox2*-C*), including ORF376 (*R. oryzae*) and ORF342 (*M. verticillata*) (Figure 3). ORF342 has no significant similarity to known endonucleases, but ORF376, like the mobile element endonuclease of *A. macrogynus* (36), encodes a protein related to homing endonucleases of the GIY-YIG type (38, 39). The same structural organization is seen in the *atp6-ORF360-atp6*-C* gene region in *A. macrogynus*, *atp9-ORF376-atp9*-C* in *R. oryzae* and *cox2-ORF342-cox2*-C* in *M. verticillata* (Figure 3), indicating the presence of similar mobile elements. However, contrary to the *A. macrogynus* case, we currently do not have biochemical evidence for the endonucleolytic activity. Both the *cox2*-C* and *atp9**-*C* fragments of the two zygomycetes encode C-terminal ends that are 100% identical at the amino acid level, suggesting that the source of potential transfers are closely related zygomycete species.

Lateral transfer of a group I intron from zygomycete to angiosperm mitochondria

A significant portion of the genomes described here is occupied by introns (*R. oryzae* 15.8%; *M. verticillata* 8.6%; and *S. culisetae* 27.4%). With 14 introns, the mtDNA of *S. culisetae* contains the largest number, 9 of which are located in the *cox1* gene (Table 1). Here, all the identified zygomycete introns are of group I and 22 contain intronic ORFs: sixteen of the LAGLI-DADG type and six of the GIY-YIG type. In *R. oryzae*, we identified one intron [*cox1*-i1(ORF305)], which is most similar to introns inserted at the same positions of angiosperm *cox1* genes (highest BLAST expect value of e^{-114} with *Philodendron oxycardium*, *Lamium* sp. and *Malpighia glabra*). Because this is the only group I intron in vascular plant mtDNAs, it has most likely been acquired by lateral transfer. The hypothesis of recent horizontal transfer of this intron from a fungal donor to flowering plants (40) is further based on the incongruence between the intron and organismal phylogenies, and its closer phylogenetic relationship to a fungal intron than to those of *Marchantia* and *Prototheca* (40-43). However, because of the high mobility of this intron and the possibility of multiple lateral transfers, the published phylogenetic inferences have to be interpreted with caution.

The presence of additional, highly similar introns in three fungi, *R. oryzae*, *S. culisetae* (this paper) and *Monoblepharella*15 (8), allows us to more rigorously test the hypothesis of Vaughn *et al*. by phylogenetic analysis (Figure 5B). Our analysis reveals that ORFs from *R. oryzae*, *Monoblepharella*15 and the three angiosperms group together with high support (98%). ORF305 from *R. oryzae* is the closest and most similar relative of the angiosperm ORFs, suggesting that the fungal donor of the group I intron and its resident ORF was a zygomycete. This scenario is biologically meaningful, because symbiotic mycorrhizal zygomycetes live in close association with most plants. However, note that ORF248 from *S. culisetae* branches with ORF313 from *P. anserina*. This observation, along with the presence of a UGA codon specifying tryptophan in ORF248 (see above), suggests a second case of lateral transfer, this time from a euascomycete close to *P. anserina*, into a zygomycete.

Phylogenetic analysis with standard mitochondrial proteins: are zygomycetes paraphyletic?

The availability of complete mtDNAs from three distant zygomycetes provides an opportunity for testing the monophyly of Zygomycota. Tree topologies were inferred from a 2890 position concatenated protein alignment using two standard ML-based methods, MrBayes and separate ML analysis (45). The results of these analyses are summarized in Figure 5A, which presents the tree produced by IQPNNI, along with bootstrap values from all methods. All methods produced similar results, except that separate analysis recovered monophyly of *S. culisetae* and *M. verticillata* (81% bootstrap support). The latter topology was also found in the 'credible set' in three independent runs of MrBayes (5.3, 44.7 and 99.0% posterior probability, respectively; MrBayes bootstrap support for this grouping was 45%). Although these data support both of these topologies, it is interesting that the monophyly of *S. culisetae* and *M. verticillata* is better supported under more sophisticated (yet statistically valid) models.

Although this phylogeny is generally robust, bootstrap support values indicate two major areas of uncertainty: the position of *A. macrogynus* and the relative branching positions of the zygomycetes. Indeed, also the approximately unbiased likelihood ratio test suggests these same problems. When sitewise likelihoods are calculated separately for each functional class, the confidence set contained a total of 17 tree topologies that failed to reject the data ($P < 0.05$). Among these topologies, *A. macrogynus* branches immediately above, immediately below or monophyletic with the chytridiomycetes, and *R. oryzae* and *M. verticillata* are found to be either monophyletic or paraphyletic (with either species branching more deeply than the other). Although *S. culisetae* and *M. verticillata* are monophyletic in the best tree under this model of separately optimized functional classes, all topologies in which all three zygomycetes are monophyletic are rejected. It is worth noting that several obvious wrong positions for *S. culisetae* were also observed (e.g. as most ancestral among fungi, or among the chytridiomycetes), suggesting that the accelerated evolutionary rate in this species causes a long-branch attraction artefact (44). Similar results are obtained when sitewise likelihoods were calculated from the fully concatenated alignment. Clearly, these data are insufficient to resolve the phylogeny of the zygomycetes, most likely because these three species diverge deeply within fungi, and at relatively short distance from each other. In such a situation, two strategies can be used to resolve the dilemma, addition of more zygomycete and neighboring fungal lineages or addition of more sequence per species. As our protein dataset is already based on complete mtDNAs, this latter strategy implies resorting to expressed sequence tag and/or nuclear genome sequences, a currently ongoing project.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Gertraud Burger (Université de Montréal) for helpful comments on the manuscript, Michael W. Gray and Murray Schnare for help in the analysis of the genes for small and large rRNAs, Kerry O'Donnell (National Center for Agricultural Utilization Research, Peoria, IL), R. W. Lichtwardt (Department of Botany, University of Kansas, Laurence, KS) and Carolyn Babcock (Canadian Collection of Fungal Cultures, Ottawa) for supply of fungal strains. We further acknowledge the excellent technical assistance of Z. Wang. Salary and interaction support from the Canadian Institutes of Health Research (MOP 42475; B.F.L.), the Canadian Institute for Advanced Research (CIAR; B.F.L.), the Nova Scotia Health Research Foundation (J.L.), and the ‘Bourses d'Excellence biT’ (CIHR;

Y.L.) and supply of laboratory equipment and informatics infrastructure by Genome Quebec/Canada is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the CIHR.

Footnotes

DDBJ/EMBL/GenBank accession nos

AY863211, AY863212, AY8632133, AY861439-AY861442

REFERENCES

1. Voigt, K., Cigelnik, E., O'Donnell, K. (1999) Phylogeny and PCR identification of clinically important Zygomycetes based on nuclear ribosomal–DNA sequence data *J. Clin. Microbiol.*, **37**, 3957–3964 .
2. Tanabe, Y., O'Donnell, K., Saikawa, M., Sugiyama, J. (2000) Molecular phylogeny of parasitic Zygomycota (Dimargaritales, zoopagales) based on nuclear small subunit ribosomal DNA sequences *Mol. Phylogenet. Evol.*, **16**, 253–262.
3. Voigt, K. and Wostemeyer, J. (2000) Reliable amplification of actin genes facilitates deep-level phylogeny *Microbiol. Res.*, **155**, 179–195.
4. Berbee, M.L., Carmean, D.A., Winka, K. (2000) Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough? *Mol. Phylogenet. Evol.*, **17**, 337–344.
5. 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.

6. James, T.Y., Porter, D., Leander, C.A., Vilgalys, R., Longcore, J.E. (2000) Molecular phylogenetics of the Chytridiomycota supports the utility of ultrastructural data in chytrid systematics *Can. J. Bot.*, **78**, 226–350.
7. Leigh, J., Seif, E., Rodriguez, N., Jacob, Y., Lang, B.F. (2003) Fungal evolution meets fungal genomics In Arora, D.K. (Ed.). *Handbook of Fungal Biotechnology*, 2nd edn NY Marcel Dekker Inc. pp. 145–161 .
8. Bullerwell, C.E., Forget, L., Lang, B.F. (2003) Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences *Nucleic Acids Res.*, **31**, 1614–1623.
9. Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., Burger, G. (2002) The closest unicellular relatives of animals *Curr. Biol.*, **12**, 1773–1778.
10. Alexopolous, C.J., Mims, C.W., Blackwell, M. *Introductory Mycology*, (1996) NY John Wiley & Sons.
11. Ustinova, I., Krienitz, L., Huss, V.A. (2000) *Hyaloraphidium curvatum* is not a green alga, but a lower fungus; A*moebidium parasiticum* is not a fungus, but a member of the DRIPs *Protist*, **151**, pp. 253–262.
12. Seif, E.R., Forget, L., Martin, N.C., Lang, B.F. (2003) Mitochondrial RNase P RNAs in ascomycete fungi: lineage-specific variations in RNA secondary structure *RNA*, **9**, 1073–1083.
13. Okpodu, C.M., Robertson, D., Boss, W.F., Togasaki, R.K., Surzycki, S.J. (1994) Rapid isolation of nuclei from carrot suspension culture cells using a BioNebulizer *BioTechniques*, **16**, 154–159.
14. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity *BMC Bioinformatics*, **5**, 113.
15. Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis *Mol. Biol. Evol.*, **17**, 540–552.
16. Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood *Syst. Biol.*, **52**, 696–704.

17. Vinh le, S. and Von Haeseler, A. (2004) IQPNNI: moving fast through tree space and stopping in time *Mol. Biol. Evol.*, **21**, 1565–1571.

18. Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., et al. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba Proc. Natl Acad. Sci. USA*, **99**, 1414–1419.

19. Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models *Bioinformatics*, **19**, 1572–1574.

20. Schmidt, H.A., Strimmer, K., Vingron, M., von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing *Bioinformatics*, **18**, 502–504.

21. Shimodaira, H. and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection *Bioinformatics*, **17**, 1246–1247.

22. Bendich, A.J. (1996) Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis *J. Mol. Biol.*, **255**, 564–588.

23. Bullerwell, C.E., Burger, G., Lang, B.F. (2000) A novel motif for identifying *rps3* homologs in fungal mitochondrial genomes *Trends Biochem. Sci.*, **25**, 363–365.

24. Lang, B.F., Gray, M.W., Burger, G. (1999) Mitochondrial genome evolution and the origin of eukaryotes *Annu. Rev. Genet.*, **33**, 351–397.

25. Bullerwell, C.E., Leigh, J., Seif, E., Longcore, J.E., Lang, B.F. (2003) Evolution of the fungi and their mitochondrial genomes In Arora, D.K. and Khachatourians, G.G. (Eds.). *Applied Mycology and Biotechnology*, Amsterdam Elsevier Science Vol. **3**, pp. 133–159.

26. Paquin, B., Laforest, M.J., Forget, L., Roewer, I., Wang, Z., Longcore, J., Lang, B.F. (1997) The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression *Curr. Genet.*, **31**, 380–395.

27. Bullerwell, C.E., Leigh, J., Forget, L., Lang, B.F. (2003) A comparison of three fission yeast mitochondrial genomes *Nucleic Acids Res.*, **31**, 759–768.

28. Turmel, M., Lemieux, C., Burger, G., Lang, B.F., Otis, C., Plante, I., Gray, M.W. (1999) The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae *Plant Cell*, **11**, 1717–1730.

29. Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., Gray, M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature *Nature*, **387**, 493–497.

30. Wise, C.A. and Martin, N.C. (1991) Dramatic size variation of yeast mitochondrial RNAs suggests that RNase P RNAs can be quite small *J. Biol. Chem.*, **266**, 19154–19157.

31. Brown, J.W. (1999) The Ribonuclease P Database *Nucleic Acids Res.*, **27**, 314.

32. Siegel, R.W., Banta, A.B., Haas, E.S., Brown, J.W., Pace, N.R. (1996) *Mycoplasma fermentans* simplifies our view of the catalytic core of ribonuclease P RNA *RNA*, **2**, 452–462.

33. Paquin, B., Laforest, M.J., Lang, B.F. (2000) Double-hairpin elements in the mitochondrial DNA of *Allomyces*: evidence for mobility *Mol. Biol. Evol.*, **17**, 1760–1768.

34. Vioque, A. (1997) The RNase P RNA from cyanobacteria: short tandemly repeated repetitive (STRR) sequences are present within the RNase P RNA gene in heterocyst-forming cyanobacteria *Nucleic Acids Res.*, **25**, 3471–3477.

35. Wolin, S.L. and Cedervall, T. (2002) The La protein *Annu. Rev. Biochem.*, **71**, 375–403.

36. Paquin, B., Laforest, M.J., Lang, B.F. (1994) Interspecific transfer of mitochondrial genes in fungi and creation of a homologous hybrid gene *Proc. Natl Acad. Sci. USA*, **91**, 11807–11810.

37. Eddy, S.R. and Gold, L. (1992) Artificial mobile DNA element constructed from the EcoRI endonuclease gene *Proc. Natl Acad. Sci. USA*, **89**, 1544–1547.

38. Burger, G. and Werner, S. (1985) The mitochondrial *URF1* gene in *Neurospora crassa* has an intron that contains a novel type of URF *J. Mol. Biol.*, **186**, 231–242.

39. Michel, F. and Cummings, D.J. (1985) Analysis of class I introns in a mitochondrial plasmid associated with senescence of *Podospora anserina* reveals extraordinary resemblance to the *Tetrahymena* ribosomal intron *Curr. Genet.*, **10**, 69–79.

40. Vaughn, J.C., Mason, M.T., Sper-Whitis, G.L., Kuhlman, P., Palmer, J.D. (1995) Fungal origin by horizontal transfer of a plant mitochondrial group I intron in the chimeric *CoxI* gene of *Peperomia J. Mol. Evol.*, **41**, 563–572.

41. Adams, K.L., Clements, M.J., Vaughn, J.C. (1998) The *Peperomia* mitochondrial *coxI* group I intron: timing of horizontal transfer and subsequent evolution of the intron *J. Mol. Evol.*, **46**, 689–696.

42. Cho, Y., Qiu, Y.L., Kuhlman, P., Palmer, J.D. (1998) Explosive invasion of plant mitochondria by a group I intron *Proc. Natl Acad. Sci. USA*, **95**, 14244–14249.

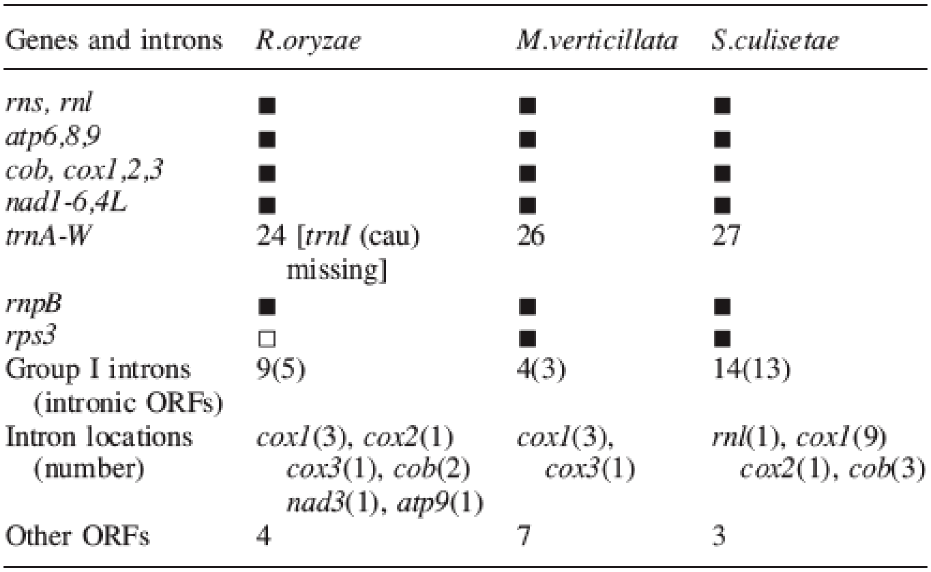
43. Palmer, J.D., Adams, K.L., Cho, Y., Parkinson, C.L., Qiu, Y.L., Song, K. (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates *Proc. Natl Acad. Sci. USA*, **97**, 6960–6966.

44. Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading *Syst. Zool.*, **27**, 401–410.

45. Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A **99**:1414-1419.

Tables

Table 1 Overview of gene, ORF and intron content in zygomycete mtDNAs



Genes and introns	*R.oryzae*	*M.verticillata*	*S.culisetae*
rns, rnl	■	■	■
atp6,8,9	■	■	■
cob, cox1,2,3	■	■	■
nad1-6,4L	■	■	■
trnA-W	24 [*trnI* (cau) missing]	26	27
rnpB	■	■	■
rps3	□	■	■
Group I introns (intronic ORFs)	9(5)	4(3)	14(13)
Intron locations (number)	*cox1*(3), *cox2*(1) *cox3*(1), *cob*(2) *nad3*(1), *atp9*(1)	*cox1*(3), *cox3*(1)	*rnl*(1), *cox1*(9) *cox2*(1), *cob*(3)
Other ORFs	4	7	3

Figure legends

Figure 1 Genomic maps of the mtDNAs of *R. oryzae*, *M. verticillata* and *S. culisetae*. The inner circle gives a scale in kilo base pair. The outer circle indicates the location of genes, exons (black) and introns plus intronic ORFs (gray). Names of ORFs, *rps3* and *rnpB* are colored to distinguish them from standard fungal genes (black).

Figure 2 Secondary structure models for mtP-RNAs from *R. oryzae*, *R. stolonifer* 194667, *R. oligosporus*, *R. spectabilis*, *M. mucedo*, *S. culisetae* and *M. verticillata*. Positions in red are invariant in the minimum bacterial consensus (32); uppercase letters in the mtP-RNAs indicate 100%, lowercase 90%, conservation of the minimum bacterial consensus sequence. The arrows pinpoint experimentally determined termini; arrow length is proportional to the percentage of molecules ending at a defined position. Double hairpin elements are named in green. The few nucleotides colored blue in the *R. stolonifer* mtP-RNA model are different in its close relative *R. oryzae*.

Figure 3 3' RNA processing motifs in zygomycetes, basidiomycetes and fission yeasts. The 3' termini of the *R.oryzae* mitochondrial SSU rRNA and of *cox1* mRNA were determined by nuclease S1 assays and run on a sequencing gel against a commercial 10 bp ladder (Invitrogen 10821-015) that was 3' labeled with ddATP (^{32}P). For experimental details see Materials and Methods. The positions of 3' termini for both molecules are indicated in the derived consensus sequences by arrows. A small fraction of the undigested form of the

SSU rRNA probe is apparent on the gel. In the lower part of the figure, additional, similar motifs in fissions yeasts and basidiomycetes are presented. Uppercase letters indicate 100% conservation and lowercase letters correspond to at least 60% nucleotide conservation. Lowercase Cs between brackets indicate the C-clusters of variable length.

Figure 4 Schematic view of atp6 regions of *A. macrogynus* and *A. arbusculus* (36), atp9 of *R. oryzae* and cox2 of *M. verticillata*. Coding sequences are enclosed in boxes and intergenic spacers are represented by a thick line. Black boxes indicate sequences present before the invasion by the corresponding ORF. Gray boxes represent ORFs and newly acquired sequences.

Figure 5 (**A**) Fungal phylogeny based on multiple proteins. Mitochondrion-encoded protein sequences from *Harpochytrium sp. 94*, *Crinipellis perniciosa*, *Cryptococcus neoformans*, *Hypocrea jecorina*, *Amoebidium parasiticum*, *Spizellomyces punctatus*, *Yarrowia lipolytica*, *Monosiga brevicolis*, *R.oryzae*, *Rhizophydium sp. JEL136*, *Penicillium marneffei*, *Pichia canadensis*, *Cantharellus cibarius*, *Sarcophyton glaucum*, *S. culisetae*, *Monoblepharella*15, *Metridium senile*, *A. macrogynus*, *Hyaloraphidium curvatum*, *Candida albicans*, *P. anserina*, *S. commune* and *M. verticillata* were aligned, concatenated and trimmed. Phylogenies were inferred from the resulting 2890 character alignment using four different methods. Shown here is the ML tree inferred using IQPNNI, along with bootstrap support values from PHYML, IQPNNI, MrBayes and separate ML analysis, in order from top to bottom, based on 100 replicates. Nodes with 100% bootstrap support using all methods are indicated by an 'asterisk'. Clearly, both the position of *A. macrogynus* and the branching order of the zygomycetes remain unclear, although the topology is robust overall. (**B**) Phylogeny of intronic ORFs. Sequences of intronic ORFs inserted in *cox1* genes were obtained from the following species: ORF305 from *R. oryzae*, ORF248 from *S. culisetae*, ORF318 from *Monoblepharella*15 (NP_803527), ORF333 from *Schizosaccharomyces japonicus* (NP_705621), ORF317 from *S. octosporus* (NP_700369), ORF313 from *P. anserina* (NP_074934), ORF319 from *Pichia canadensis* (NP_038209), ia4 from *S. cerevisiae* (AAB21126, ORF251 *Chlorella vulgaris* (T07187), ORF234.2 *Prototheca wickerhamii* (NP_042245), ORF280 from *Peperomia obtusifolia* (AAB86934, ORF279 from *Veronica catenata* (CAA11340 and ORF277 from *Maranta leuconeura* (CAA11350. Sequences were aligned as described in Materials and Methods, and a phylogeny was inferred by ML. Only bootstrap support values >50% are shown. This tree robustly supports the monophyly between sequences from *Monoblepharella*15, *R. oryzae*, and the angiosperms, strongly suggesting horizontal intron transfer between these two groups.

Figure 1

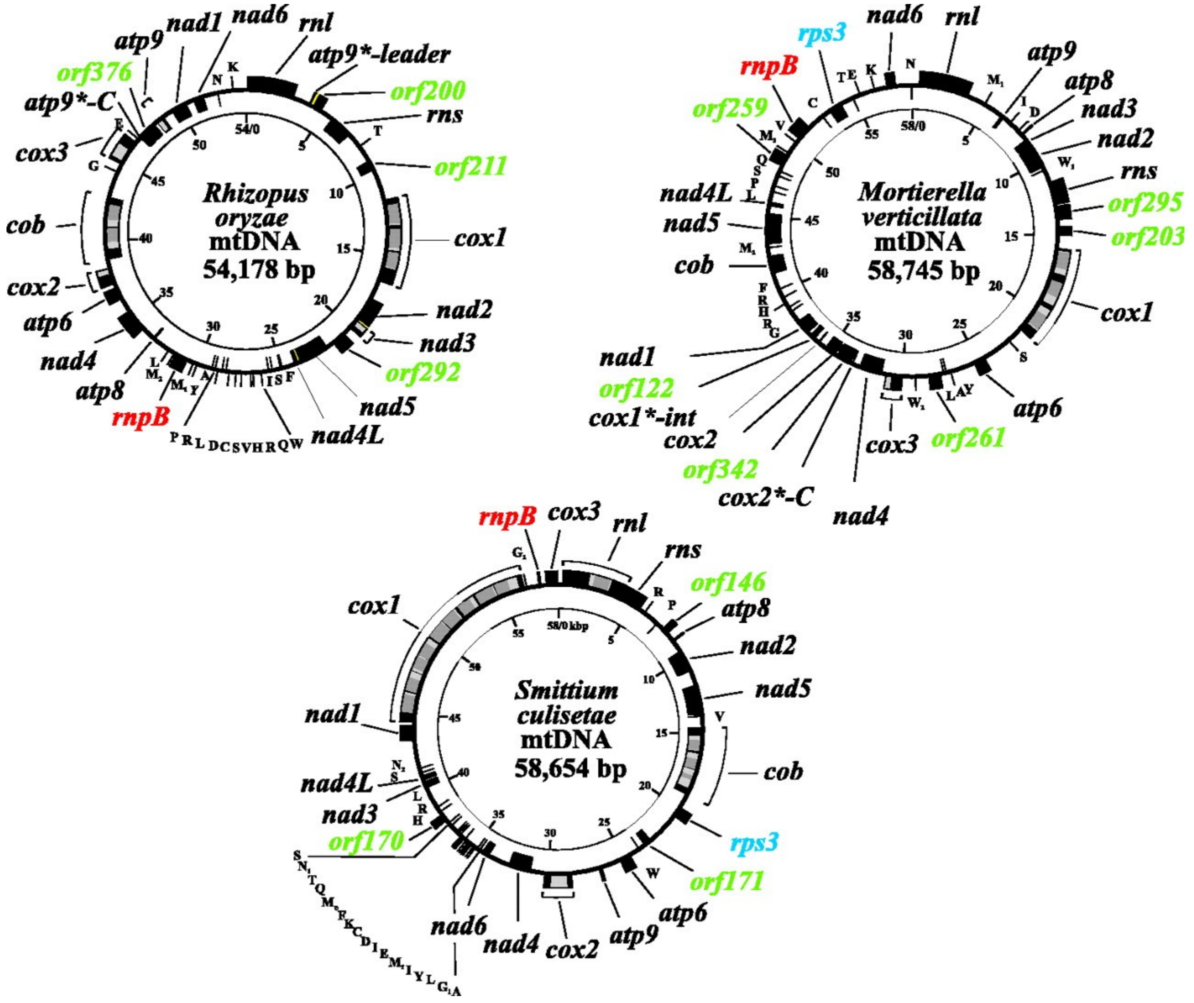

Rhizopus oryzae mtDNA 54,178 bp
54/0
5
10
15
20
25
30
35
40
45
50
nad1
nad6
rnl
atp9
atp9*-leader
orf376
orf200
atp9*-C
rns
N
K
T
E
G
cox3
orf211
cob
cox1
cox2
nad2
atp6
nad3
nad4
orf292
atp8
rnpB
nad5
nad4L
L
M2
M1
Y
A
I S F
P R L D C S V H R Q W
Mortierella verticillata mtDNA 58,745 bp
58/0
5
10
15
20
25
30
35
40
45
50
55
rps3
nad6
rnl
rnpB
atp9
atp8
nad3
nad2
rns
orf295
orf203
orf259
nad4L
nad5
cob
cox1
nad1
orf122
cox1*-int
cox2
orf342
cox2*-C
nad4
cox3
orf261
atp6
T E K N
C
V
M2
Q S P L
M1
F R H R G
M1
I D
W1
S
W2
L A Y
Smittium culisetae mtDNA 58,654 bp
58/0 kbp
5
10
15
20
25
30
35
40
45
50
55
rnpB
cox3
rnl
rns
orf146
atp8
nad2
nad5
cox1
nad1
nad4L
nad3
orf170
cob
rps3
orf171
atp6
atp9
cox2
nad4
nad6
G2
R
P
V
W
N2
S
L
R
H
S N1 T Q M3 F K C D I E M2 I Y L G1 A

Figure 2

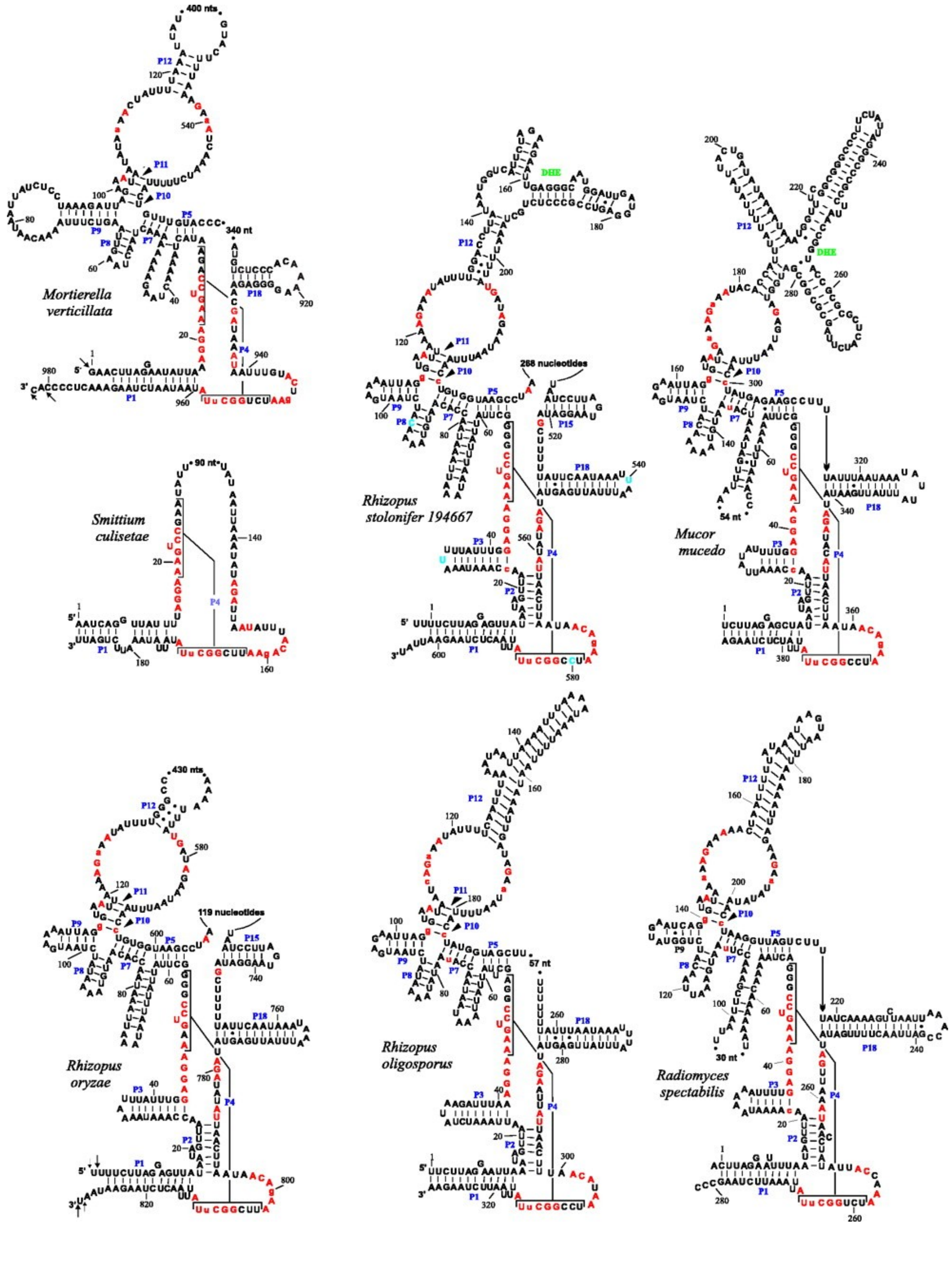
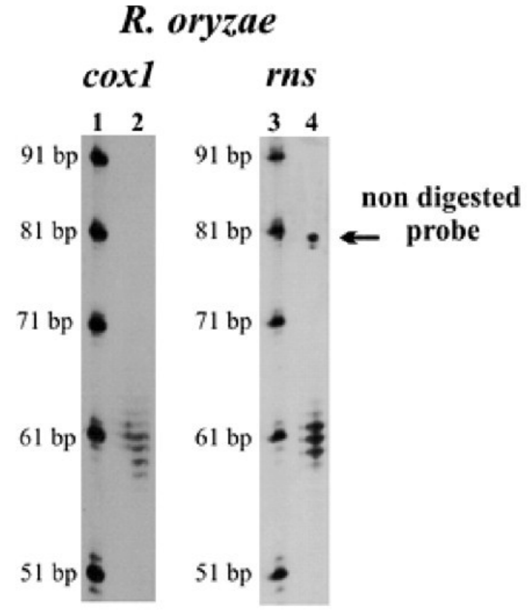

Mortierella verticillata
400 nts
340 nt
Smittium culisetae
90 nt
Rhizopus stolonifer 194667
268 nucleotides
DHE
Mucor mucedo
54 nt
Rhizopus oryzae
430 nts
119 nucleotides
Rhizopus oligosporus
57 nt
Radiomyces spectabilis
30 nt
P1
P2
P3
P4
P5
P7
P8
P9
P10
P11
P12
P15
P18
5'
3'

Figure 3



M. verticillata

```
cox1   TCATCCC..TCT
cox2   ACTCCCCC.TTT
cox3   ACACCCCC.TTC
cob    ACACCCCCCTTA
nad1   ACACCCCCCTAA
nad6   ACACCCCCCATC
atp6   ACACCCCC.TTA
atp9   ACACCCCCCTTT
rns    AATCCCCC.TTA
              ↓
rnpB   ACTCCCC..ACT
cons.  acacCC[ccc]tt
```

S. culisetae

```
cox1 TAATACCCCCC.TTAA
cox3 TAATACCCCCCCTTTT
nad1 TAATACCCC...TTTG
nad4 TAATACCCC...TTAA
nad5 TAATACCCCCC.TTAA
nad6 TAATACCCCC..TTTA
atp6 TAATACCCCC..TTAA
atp8 TAATACCCC...TTTT
atp9 TAATACCCC...TTTT
rps3 TAATACCCCCC.TTAA
rns  TAATACCCCCC.TTAT
cons.TAATACCCC[ccc]TT
```

R. oryzae

```
              ↓     ↓
cox1   ACACCCCC.....TTAGA
cox2   CCACCCCCTCCCCTTACT
cox3   CCACCC.......AAAAA
cob    CCACCC.......TTATT
nad4   TTACCCCCC....TTTTA
nad5   ATTCCCCCC....TCAAT
atp9   GGACCCACC....TTAGA
                ↓   ↓↓
rns    AGTCCCTCCC...AATAT
cons.    aCCC[ccccccc]t
```

Consensus motifs

Rhizopus oryzae	**aCCC[ccccccc]t**
Mortierella verticillata	**acacCC[ccc]....tt**
Smittium culisetae	**TAATACCCC[ccc]...TT**
Schizophyllum commune	**aaTACCCC[ccccc]**
Cryptococcus neoformans	**AcTCCCC**
Crinipellis perniciosa	**AcCCCc.......A**
Schizosaccharomyces pombe	**ttCCCC[cccc]..TT**
Schizosaccharomyces octosporus	**tAACCCCC[cccc]**
Schizosaccharomyces japonicus	**tAACCCCC[cccc]**

Figure 4

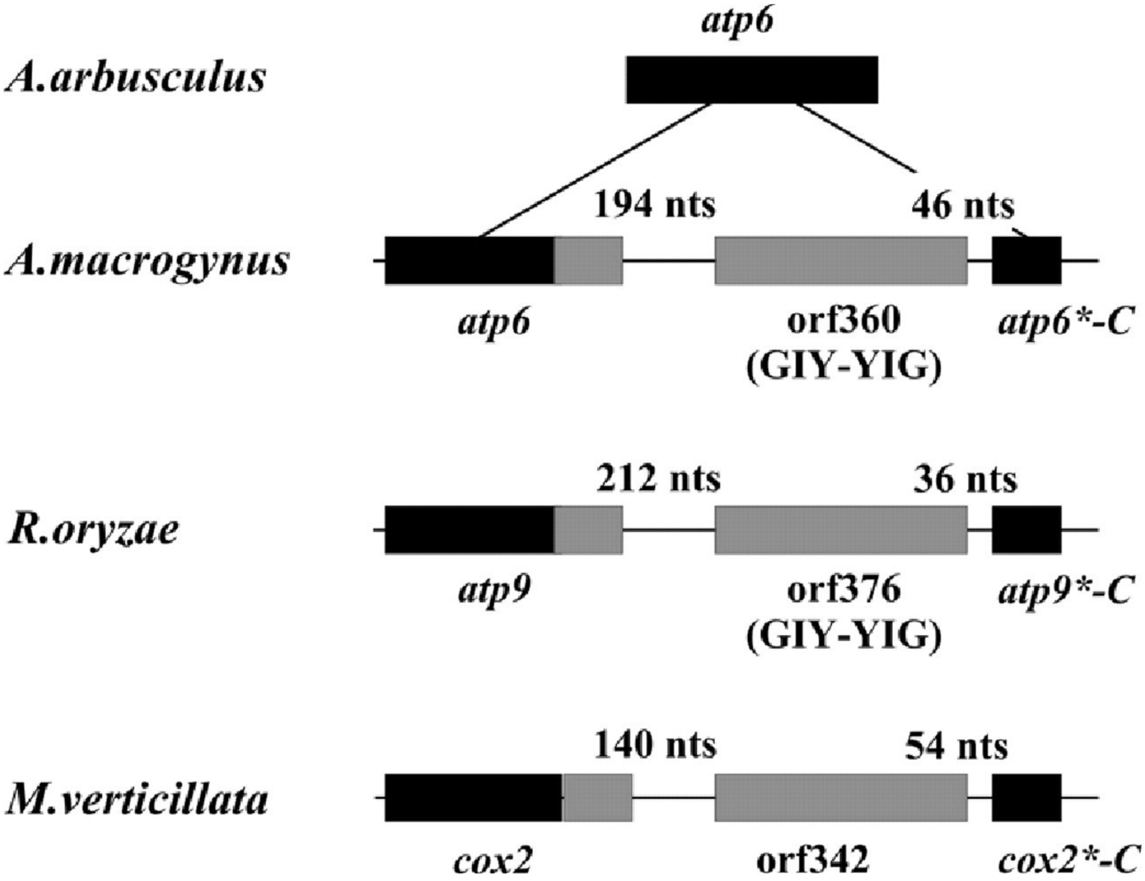
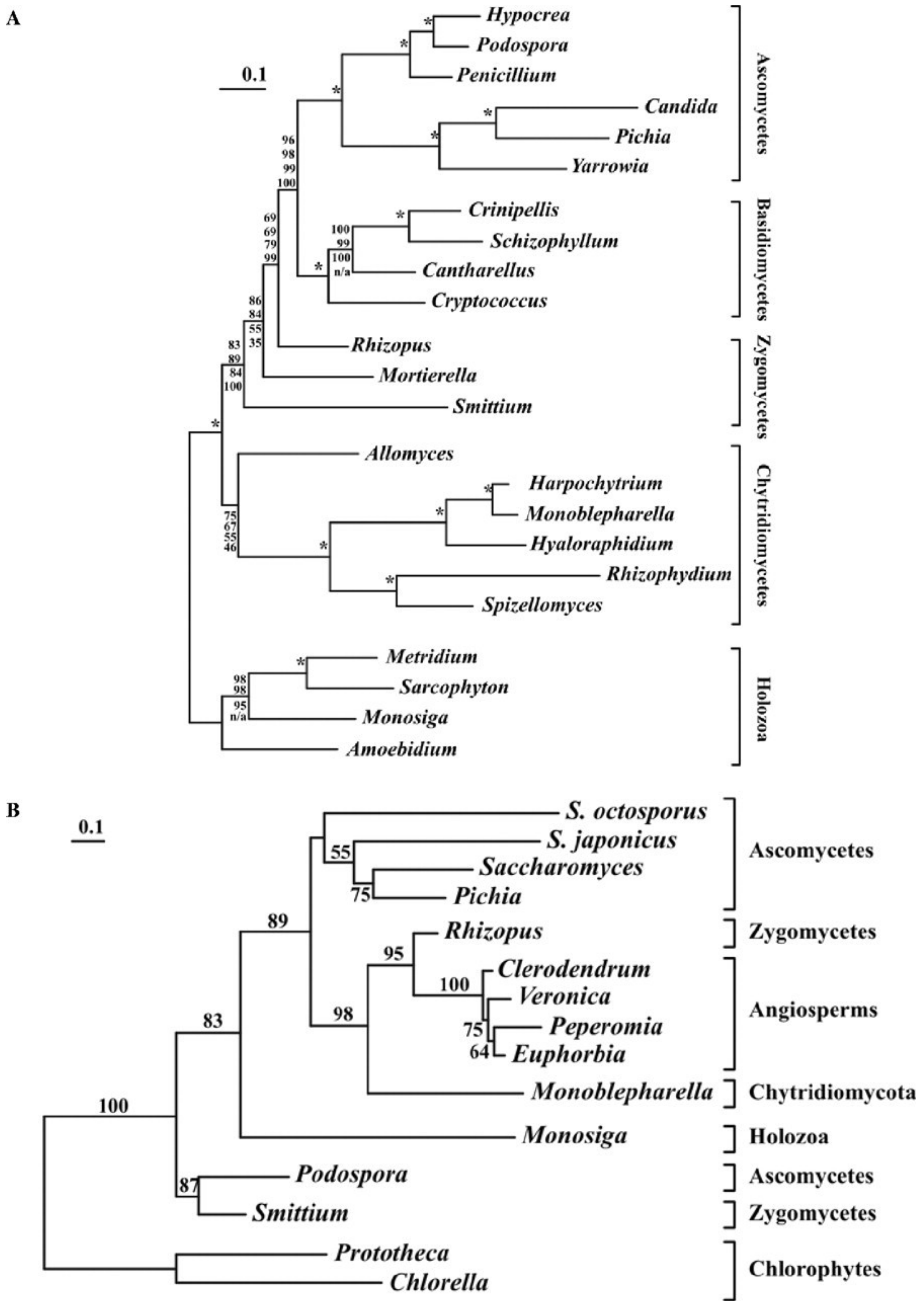

atp6
A.arbusculus
194 nts
46 nts
A.macrogynus
atp6
orf360
(GIY-YIG)
atp6*-C
212 nts
36 nts
R.oryzae
atp9
orf376
(GIY-YIG)
atp9*-C
140 nts
54 nts
M.verticillata
cox2
orf342
cox2*-C

Figure 5


A
0.1
Hypocrea
Podospora
Penicillium
Candida
Pichia
Yarrowia
Crinipellis
Schizophyllum
Cantharellus
Cryptococcus
Rhizopus
Mortierella
Smittium
Allomyces
Harpochytrium
Monoblepharella
Hyaloraphidium
Rhizophydium
Spizellomyces
Metridium
Sarcophyton
Monosiga
Amoebidium
Ascomycetes
Basidiomycetes
Zygomycetes
Chytridiomycetes
Holozoa
*
96 98 99 100
69 69 79 99
86 84 55 35
83 89 84 100
100 99 100 n/a
75 67 55 46
98 98 95 n/a
B
0.1
S. octosporus
S. japonicus
Saccharomyces
Pichia
Rhizopus
Clerodendrum
Veronica
Peperomia
Euphorbia
Monoblepharella
Monosiga
Podospora
Smittium
Prototheca
Chlorella
Ascomycetes
Zygomycetes
Angiosperms
Chytridiomycota
Holozoa
Ascomycetes
Zygomycetes
Chlorophytes
55
75
89
95
100
98
75
64
83
100
87

Supplementary Information:

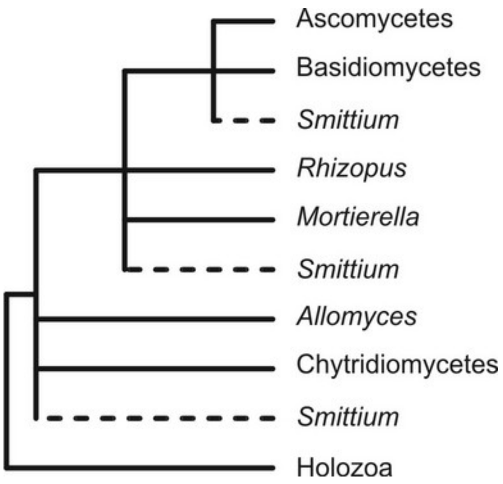


Figure 1S:

Legend to Figure 1S: Strict consensus of tree topologies examined under separate analysis and likelihood ratio tests. For these analyses, all possible topologies were generated, with certain constraints (representing groups that received a minimum of 95% bootstrap support under standard ML analysis). Ascomycetes, basidiomycetes, ascomycetes + basidiomycetes, and chytridiomycetes (excluding *A. macrogynus*) were constrained as monophyletic groups, with the Holozoa as outgroup. Additionally, only topologies in which *A. macrogynus* appeared as a monophyletic member of the chytrids, at the base of the fungal clade, or at the base of all fungi, excluding the chytrids were included. Likewise, only topologies in which *R. oryzae* or *M. verticillata* formed a monophyletic group with the ascomycetes plus basidiomycetes (to the exclusion of the chytrids) were considered. Finally, the previous rules were relaxed for *S. culisetae*, which was allowed to be positioned anywhere among the ingroup species, except within the ascomycetes, basidiomycetes, or chytridiomycetes (excluding *A. macrogynus*). For this reason, all positions considered for *S. culisetae* are indicated with a dotted line. These rules resulted in a set of 99 distinct topologies (available upon request).

Table 1S: Codon usage in the mtDNAs of *R. oryzae*, *M. verticillata*, *S. culisetae*

```
==========================================================================
F TTT 226,199,358   S TCT 150,135,150   Y TAT 141,162,254   C TGT  21, 23,36
      130,117,252          81, 84,111         111,108,272          24, 23,41

F TTC 119,184, 29   S TCC   3,  7,  1   Y TAC  35, 40,  4   C TGC  --, --,--
       25, 20, 14           9, 11,  9           24, 17, 6            5, --, 2

L TTA 542,566,592   S TCA  90,132,108   * TAA  13, 10, 15   * TGA  --,  2,--
      200,211,424          54, 62, 60            7,  8, 14           --, --, 2

L TTG   2, 13,  1   S TCG   6,  6,  2   * TAG   1,  3, --   W TGG  77, 62,57
       29, 14, 22           9,  8,  5            3,  1,  2          31, 21,39
==========================================================================
L CTT  62, 65, 41   P CCT  94, 87,121   H CAT  72, 74, 72   R CGT  --,  6, 7
       56, 45, 26          45, 34, 60           32, 37, 60            8, 16,20

L CTC  --,  1,  1   P CCC   2,  7,  3   H CAC  15, 10,  2   R CGC  --, --,--
        3,  3,  5           4, --,  3            6,  3,  3            3, --,--

L CTA  60, 21, 12   P CCA  71, 87, 40   Q CAA  79, 88, 58   R CGA  46, 19, 1
       36, 20, 24          24, 23, 23           81, 47,109           19, 10, 6

L CTG   2,  3, --   P CCG  --,  7,  2   Q CAG   4,  2, --   R CGG  --,  3, 1
        3,  5,  4           7,  3,  2            5,  6,  3            6,  2, 1
==========================================================================
I ATT 342,249,259   T ACT 131,131, 97   N AAT 146,155,237   S AGT  89, 72,109
      222,119,228          82, 68, 98         184,204,436          56, 45, 80

I ATC  31, 91,  4   T ACC   1,  4,  6   N AAC  26, 55,  4   S AGC   4, 20,  1
       32, 18, 21           7,  9,  5           31, 15, 10            4, 11,  2

I ATA  --,173,433   T ACA  99,128,120   K AAA  97,137,170   R AGA  38, 66, 55
       17,121,266          48, 50, 60         224,210,501          68, 42, 99

M ATG 129,117, 91   T ACG   2,  6,  4   K AAG   3,  6, --   R AGG  --, --, --
      121, 32, 48           4,  3,  4           37, 25, 17            5,  3,  5
==========================================================================
V GTT 117, 86, 77   A GCT 167,177,113   D GAT  90,102, 92   G GGT 133,154,198
       37, 41, 44          47, 62, 59        116, 82,156          54, 54,126

V GTC  --,-- ,  1   A GCC   5, 17,  3   D GAC   4,  4,  1   G GGC  --, --, --
        6,  2,  3          10,  4,  1          9,  6,  6           6,  3,  2

V GTA 217,194, 99   A GCA 116,103, 72   E GAA 103, 98, 97   G GGA 171,135,58
       69, 41, 54          31, 30, 27        115, 97,166          51, 45,55

V GTG   4, 12,  4   A GCG  14, 15,  2   E GAG  15, 18,  2   G GGG   6, 24, 5
        7,  7,  1           5,  2,  2         16, 18,  7           7,  9, 6
=============================================================================
```

Cognate amino acids are specified in the one-letter code (asterisks = stop codon). The upper rows of numbers indicate the total number of codons for standard protein-coding genes of *R. oryzae*, *M. verticillata*, and *S. culisetae*. Lower numbers specify the total number of codons in intronic ORFs of the three species, respectively.

Chapter 3 Phylogenomic analyses support the monophyly of Taphrinomycotina, including *Schizosaccharomyces* fission yeasts

Yu Liu[1], Jessica W. Leigh[2], Henner Brinkmann[1], Melanie T. Cushion[3], Naiara Rodriguez-Ezpeleta[1], Hervé Philippe[1], and B. Franz Lang[1]*

[1]*Robert Cedergren Centre, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, C.P. 6128, Montréal (Québec), H3T 1J4, Canada*

[2]*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, (Nova Scotia) B3H 4H7, Canada*

[3]*Department of Internal Medicine, Division of Infectious Diseases, University of Cincinnati College of Medicine, Cincinnati, Ohio 45220, US A*

Corresponding author: B. Franz Lang

Running head: Monophyly of Taphrinomycotina

Keywords: Fungi, Taphrinomycotina, *Schizosaccharomyces*, *Pneumocystis*, phylogeny, mitochondria, long-branch attraction artifact

Summary

Several morphologically dissimilar ascomycete fungi including *Schizosaccharomyces*, *Taphrina*, *Saitoella*, *Pneumocystis* and *Neolecta* have been grouped into the taxon Taphrinomycotina (Archiascomycota or Archiascomycotina), originally based on rRNA phylogeny. These analyses lack statistically significant support for the monophyly of this grouping, and although confirmed by more recent multi-gene analyses, this topology is contradicted by mitochondrial phylogenies. To resolve this inconsistency, we have assembled phylogenomic mitochondrial and nuclear datasets from four distantly related taphrinomycotina taxa: *Schizosaccharomyces pombe*, *Pneumocystis carinii*, *Saitoella complicata* and *Taphrina deformans*. Our phylogenomic analyses based on nuclear data (113 proteins) conclusively support the monophyly of Taphrinomycotina, diverging at the base of Ascomycota. However, despite the improved taxon sampling, Taphrinomycotina continue to be paraphyletic with the mitochondrial dataset (13 proteins): *Schizosaccharomyces* species associate with budding yeasts (Saccharomycotina), and the other Taphrinomycotina group at the base of Ascomycota. Yet, as *Schizosaccharomyces* and Saccharomycotina species are fast-evolving, the mitochondrial phylogeny may be influenced by a long-branch attraction (LBA) artifact. After removal of fast-evolving sequence positions from the mitochondrial dataset, we recover the monophyly of Taphrinomycotina. Our combined results suggest that Taphrinomycotina is a legitimate taxon, that this group of species diverges at the base of Ascomycota, and that phylogenetic

positioning of yeasts and fission yeasts with mitochondrial data is plagued by a strong LBA artifact.

Introduction

Ascomycota are currently subdivided into three major taxa (Hibbett et al. 2007): Saccharomycotina (Hemiascomycota; budding yeasts), Pezizomycotina (Euascomycota; for the most part filamentous fungi, e.g. *Neurospora*) and Taphrinomycotina (Archiascomycota). The taxon Taphrinomycotina was initially created based on rRNA phylogeny (Nishida and Sugiyama 1993), regrouping diverse fungal species of previously uncertain taxonomic affiliation: (i) *Schizosaccharomyces* species (fission yeasts; previously considered to be highly divergent yeast lineages), (ii) *Taphrina* (several fungal plant pathogens), (iii) the anamorphic yeast-like *Saitoella*, a suspected ascomycete or basidiomycete, and (iv) *Neolecta irregularis*, a fungus with filamentous cell growth that forms complex fruiting bodies (unique within this group of organisms). Yet, the statistical support for this grouping with rRNA data is well below standards (for details, see (Leigh et al. 2003)). Addition of potential taphrinomycotina taxa, for instance more *Schizosaccharomyces* species or *Pneumocystis carinii* (a unicellular lung pathogen (Edman et al. 1988b) that like *Schizosaccharomyces* divides by binary fission), has not improved the outcome. Evidently, resolving this question requires substantially more than just rRNA data.

Several multi-gene analyses have more recently been conducted to overcome the apparent problems with inferring fungal relationships. These analyses differ in their choice of genes. First, datasets with six or fewer nuclear genes also produce conflicting phylogenies. For instance in an early overview paper (Baldauf et al. 2000), *Schizosaccharomyces*, the only Taphrinomycotina included in this analysis, groups with Saccharomycotina, although without significant support. This topology is recovered by a more recent analysis (Diezmann et al. 2004), but contradicted by others (James et al. 2006a; Liu, Hodson, and Hall 2006; Spatafora et al. 2006; Sugiyama, Hosaka, and Suh 2006) that find high bootstrap support for Taphrinomycotina as a monophyletic grouping at the base of Ascomycota. Yet, rigorous statistical testing (e.g., by applying the AU test (Shimodaira 2002)) has not been performed in these cases, and because most sequence information was obtained by PCR, only limited genomic sequence information was available to exclude potentially misleading gene paralogs with confidence. Additional reasons why analyses with small datasets are more likely misled by phylogenetic artifacts are discussed elsewhere (Delsuc, Brinkmann, and Philippe 2005). Finally, in two of these analyses (James et al. 2006a; Spatafora et al. 2006), both rRNA and protein sequences were used in the same dataset, which implies the use of potentially problematic mixed-model analyses that preclude rigorous statistical AU testing. The applied Bayesian analyses are known to largely overestimate confidence when using real data, as these evolve in much more complex ways than implemented in current models (Erixon et al. 2003; Taylor and Piel 2004; Mar, Harlow, and Ragan 2005). In turn, when applying the AU test to alternative analyses that are restricted to the nucleotide level, the risk of error due to compositional bias (rRNA *versus* protein gene sequences) increases.

In phylogenomic analyses that do not suffer from lack of sequence data, *S. pombe* consistently diverges at the base of Ascomycota with significant support (e.g., (Philippe et al. 2004; Fitzpatrick et al. 2006; Robbertse et al. 2006; Dutilh et al. 2007)). Yet, the question of Taphrinomycotina monophyly remains open, as genome-size datasets are not available for other taphrinomycotina lineages. Finally, mitochondrial datasets with 13 proteins and three *Schizosaccharomyces* species consistently support a grouping of fission yeasts with Saccharomycotina (Bullerwell, Forget, and Lang 2003a; Leigh et al. 2003). Obviously, the use of multi-gene datasets is insufficient to tackle the given phylogenetic question without paying close attention to potential phylogenetic artifacts (Delsuc, Brinkmann, and Philippe 2005). In the analyses of mitochondrial data (Bullerwell, Forget, and Lang 2003a; Leigh et al. 2003), the authors suggest that the grouping of Saccharomycotina and *Schizosaccharomyces* may be due to an LBA artifact, which causes clustering of fast-evolving lineages irrespective of their true evolutionary relationships. A common strategy to overcome this artifact involves the complete elimination of fast-evolving species; yet in the mitochondrial dataset, all *Schizosaccharomyces* and budding yeast species are fast-evolving. Other less radical options include the exclusion of fast-evolving sequence positions (Brinkmann and Philippe 1999), or the use of more realistic evolutionary models, e.g. the CAT model (Lartillot, Brinkmann, and Philippe 2007). Evidently, such improvements at the analytical level should be combined with improved taxon sampling, with particular emphasis on the addition of slowly-evolving species. Finally, congruence with analyses with alternative datasets (e.g., nuclear *versus* mitochondrial) is an indication that results are accurate.

In the present study, we take advantage of new data provided by both nuclear and mitochondrial genome projects for all key taphrinomycotina species except *Neolecta*, which unfortunately has not yet been grown in culture. We compare two large datasets, one with 113 nuclear and another with 13 mitochondrial proteins, and conclude that Taphrinomycotina is indeed a monophyletic group diverging at the base of Ascomycota.

Material and Methods

Construction of cDNA libraries and EST sequencing

Saitoella complicata (NRLL Y-17804) and *Taphrina deformans* (NRRL T-857) cDNA libraries were constructed from strains grown on glycerol medium, following recently published protocols (Rodriguez-Ezpeleta et al. in press). Plasmids were purified using the QIAprep 96 Turbo Miniprep Kit (Qiagen), sequencing reactions were performed with the ABI Prism BigDyeTM Terminators version 3.0/3.1 (Perkin-Elmer, Wellesley, MA, USA) and a total of 3840 *S. complicata* and 3919 *T. deformans* ESTs were sequenced on an MJ BaseStation. Trace files were imported into the TBestDB database (http://tbestdb.bcm.umontreal.ca/searches/login.php) (O'Brien et al. 2007) for automated processing, including assembly as well as automated gene annotation by AutoFact (Koski et al. 2005a). *P. carinii* sequences were obtained from the *Pneumocystis* Genome Project (http://pgp.cchmc.org).

Mitochondrial sequencing

S. complicata and *T. deformans* were grown with vigorous shaking in liquid medium (1% yeast extract, 3% glycerol). The harvested cells were disrupted by manual shaking with glass beads, and mitochondrial DNA was isolated following a whole cell lysate protocol (Lang and Burger 2007), and sequenced using a random procedure (Burger et al. 2007).

Dataset construction

The nuclear dataset was assembled by adding EST and genomic sequences from GenBank to a previously published alignment (Rodriguez-Ezpeleta et al. 2007a). Paralogous proteins were identified and removed from the alignment as described (Roure, Rodriguez-Ezpeleta, and Philippe 2007). Gblocks (Castresana 2000) (default parameters) was used to extract unambiguously aligned regions. The inclusion of some missing data allowed us to add more genes and species. From originally 174 proteins, 113 were selected to minimize the degree of missing data in phylogenetic analysis. The final alignment has a total number of 29 387 amino acid positions and 54 species. The average proportion of missing data is 25% per species. The proportion of missing data for each species is listed in Supplementary Material (Table S1).

The mitochondrial protein alignment includes our new *T. deformans* and *S. complicata* sequences, as well as sequences retrieved from public data repositories (Genbank and the *Pneumocystis* Genome Project)http://pgp.cchmc.org/. Sequences of 13 mitochondrial proteins (*cox1, 2, 3, cob, atp6, 9,* and *nad1, 2, 3, 4, 4L, 5, 6*) were selected for phylogenetic analysis. An application developed in-house (mams) was used for automatic protein alignment with Muscle (Edgar 2004), removal of ambiguous regions with Gblocks (Castresana 2000), and concatenation. The final dataset contains 2 596 amino acid positions with missing data only in *Schizosaccharomyces* species and in *Saccharomyces cerevisiae* (46.2% missing positions for those species), which both lost all *nad* genes, coding for subunits of complex I of the respiratory chain.

Phylogenetic analysis of the nuclear dataset

Phylogenetic analyses were performed at the amino acid level. The concatenated nuclear protein datasets were analyzed either by maximum likelihood (ML) or Bayesian inference (BI) methods. Three ML programs, Treefinder (Jobb, von Haeseler, and Strimmer 2004), PhyML (Guindon and Gascuel 2003a), and RAxML (Stamatakis 2006) were used with the WAG+gamma model with four categories. In case of BI methods, we used either MrBayes ((Ronquist and Huelsenbeck 2003); WAG+gamma model, 500,000 generations, first 100,000 generations removed as burn-in, analysis repeated three times with identical results), or PhyloBayes (version 2) ((Lartillot and Philippe 2004); CAT model, 3000 cycles, first 1000 cycles removed as burn-in, analysis repeated three times with identical results)). The reliability of internal branches was either evaluated based on 100 (ML) bootstrap replicates, or on posterior probabilities.

Likelihood tests of competing tree topologies were also performed. 945 topologies were generated by constraining trusted internal branches (monophyly of Saccharomycotina, Pezizomycotina, Ascomycota, Basidiomycota, and the grouping of Zygomycota and Chytridiomycota), leaving the four Taphrinomycotina unconstrained within Ascomycota. The site-wise likelihood values for each topology were estimated using Tree-Puzzle (Schmidt et al. 2002), and p-values for each topology were calculated with CONSEL (Shimodaira and Hasegawa 2001).

Phylogenetic analysis of the mitochondrial dataset with the SF method

LBA artifacts may possibly be overcome by elimination of fast-evolving sequence positions with the slow-fast (SF) method (Brinkmann and Philippe 1999). Briefly, the dataset is split into monophyletic groups, and the number of substitutions for each position in each group is estimated using a parsimony criterion with PAUP* (Swofford 2000). These numbers are summed over all groups of the dataset, providing an estimate of the variability for each position. A number of data sets (in the current analysis, 14) are then constructed with an increasing fraction of fast-evolving sequence positions.

Trees and bootstrap support (100 replicates) for the sub-datasets were estimated with RAxML, as Treefinder and PhyML were often trapped in local optima with these relatively small datasets.

Results

Phylogenetic analysis of the nuclear dataset

Our nuclear dataset contains 113 orthologous proteins (29 387 amino acid positions) from 54 fungal species, including 33 Ascomycota and representatives of the three other major fungal groups (Basidiomycota, Zygomycota and Chytridiomycota). In the phylogenetic tree shown in Fig. 1, the monophyly of Ascomycota, Saccharomycotina, Pezizomycotinaand Basidiomycota are recovered with significant support by both ML and BI methods. In addition, Taphrinomycotina form a significantly supported monophyletic group (> 99% bootstrap proportion (BP) and 1.0 posterior probability (PP)). The grouping of *S. pombe* with *P. carinii* receives 95% support using Treefinder, 86% with RAxML and 98% with PhyML; the branching order of *S. complicata* and *T. deformans* remains unresolved.

Datasets including ESTs usually contain a fraction of missing data, amounting for *S. complicata* and *T. deformans* to 66.8% and 56.8%, respectively. The data set contains 113 proteins, but only one single protein contains sequences from all 54 species (rpl4B). To test the potential influence of missing data, we reduced the dataset to the most complete 76 proteins, thereby decreasing missing positions for these two species to 43.0% and 39.9%, respectively. The inferred tree topologies remain the same, and support values for the monophyly of Taphrinomycotina are only slightly reduced (ML inferences, BP > 95%; MrBayes, PP 1.0; PhyloBayes PP 0.99, Supplementary Fig. S1).

Likelihood test of competing topologies

Using the original complete nuclear dataset, both ML and BI approaches yield identical, well-supported tree topologies. To assess the level of confidence with a strict, alternative approach, we performed likelihood-based tests of competing tree topologies with CONSEL

(Shimodaira and Hasegawa 2001), with the complete dataset (113 proteins). The corresponding 10 top-ranking topologies according to AU test p-values are shown in Table 1. All scenarios in which Taphrinomycotina are paraphyletic are rejected with confidence ($p < 0.01$), thus confirming the monophyly of Taphrinomycotina as well as their position at the base of Ascomycota.

Phylogenetic analysis of mitochondrial datasets

The mitochondrial dataset contains 2 596 amino acid positions from 13 well-conserved mitochondrial proteins, including 29 species from the four major fungal lineages. In ML analyses, the newly added Taphrinomycotina (*T. deformans, P. carinii,* and *S. complicata*) group at the base of Ascomycota (Fig. 2), and as in previously published analyses (Bullerwell, Forget, and Lang 2003a; Leigh et al. 2003; Pramateftaki et al. 2006), *S. pombe*, *S. japonicus*, and *S. octosporus* group with Saccharomycotina. Yet, due to the addition of the new Taphrinomycotina species, the support for the grouping of fission yeasts with budding yeasts is noticeably lower (Fig. 2; 92% with Treefinder, 76% with RAxML and 87% with PhyML). In our experience, the heuristic search of RAxML is most effective in avoiding local minima; thus, the 76% confidence level of RaxML in this analysis is the most reliable. BI analyses (using MrBayes and PhyloBayes) inferred the same topology as ML approaches, with more than 0.99 PP for all internal branches except the one leading to Chytridiomycota (0.64 PP).

As *Schizosaccharomyces* and Saccharomycotina species have relatively long branches, they are suspected to group due to a LBA artifact. If this interpretation is correct, removing Saccharomycotina is expected to relocate the *Schizosaccharomyces* to its correct position. Indeed, instead of grouping with Pezizomycotina, three *Schizosaccharomyces* group with other Taphrinomycotina after Saccharomycotina are removed. The monophyly of Taphrinomycotina receives varying support (Treefinder, BP 95%, RAxML, 66%, PhyML, 97%; MrBayes, PP 1.0, Fig. 3).

We further explored the use of a fast-evolving fungal outgroup, which was expected to draw *Schizosaccharomyces* away from Saccharomycotina, to a more basal position. To test this prediction, we reduced the original mitochondrial dataset to 19 species, including all 15 Ascomycota plus four (fast-evolving) Chytridiomycota. Indeed, analyses of this dataset with ML and BI methods position *Schizosaccharomyces* at the base of Fungi (Supplementary Fig. S2), although with marginal support (Treefinder, BP 71%, RAxML, 57%, PhyML, 86% MrBayes, PP 0.95).

Finally, we analyzed the mitochondrial dataset with the SF method, which is designed to reduce the effect of LBA by selecting slowly-evolving positions, thus increasing the ratio of phylogenetic signal to noise (Delsuc et al. 2005). A series of data matrices containing increasing fractions of fast-evolving positions were analyzed with both ML and BI methods (Fig. 4). In the datasets with the most slowly evolving sites and most reliable phylogenetic information (S2-S5; only results from S3 and S5 are shown in Fig. 4A, for more details see Fig S3 in supplementary information), the *Schizosaccharomyces* lineage grouped together with other Taphrinomycotina, at the base of Ascomycota. Yet, although there was good support (BP of 96 or 88%) to reject a grouping of

Saccharomycotina plus Taphrinomycotina, there was not significant BP support to the monophyly of Taphrinomycotina. This result is most likely due to the small size of the remaining datasets (S2 contains only 1,023 amino acid positions, S3: 1,223, S4: 1,436 and S5: 1,638), and a relatively weak phylogenetic signal. In fact, addition of further fast-evolving positions to S5 resulted in decrease of support, as expected in a classical case of LBA. Finally, as more fast-evolving positions were included (the S6 - S14 datasets), *Schizosaccharomyces* grouped with Saccharomycotina, and the BP for this incorrect topology increased (the result from S7 and S9 are shown in Fig. 4B). The evolution of BP supports for all S datasets is shown in Supplementary Fig. S3.

Discussion

The nuclear dataset significantly supports the monophyly of Taphrinomycotina

Our phylogenetic analysis is first in using a large number of nucleus-encoded proteins from most key taphrinomycotina species, and concludes with high confidence that Taphrinomycotina is monophyletic, branching at the base of Ascomycota. Some authors have claimed that missing data (in our case, due to partial EST sequencing) may result in unstable tree topologies (Anderson 2001; Sanderson et al. 2003). Yet, consistent with other work (Wiens 2003; Philippe et al. 2004), our ML analysis did not confirm this claim. Our explanation is that the effect of missing data is negligible when using large datasets with a strong phylogenetic signal. The comparison of alternative topologies with the AU test confirmed the monophyly of Taphrinomycotina with high confidence ($p < 0.01$), although the relationships among Taphrinomycotina remain to be resolved. Additional data from the ongoing *S. octosporus* and *S. japonicus* genome projects are expected to improve tree resolution, and complete genome sequences from *Taphrina* and *Saitoella* (slowly-evolving taphrinomycotina genomes that we expect to be more gene-rich and more typical for Taphrinomycotina than those of *Schizosaccharomyces*) are required for confident inference of their phylogenetic position. Finally, EST or genome sequencing will be required to confirm that *Neolecta irregularis* belongs in Taphrinomycotina.

The mitochondrial tree topology is sensitive to phylogenetic artifacts

Mitochondrion-encoded protein data have been successfully used to resolve a large variety of phylogenetic questions, in some cases predicting for the first time deep relationships with high confidence (e.g., (Lang et al. 2002a)). Yet, mitochondrial genes tend to have a high A+T sequence bias which contributes to phylogenetic artifacts, particularly in lineages with elevated evolutionary rates. For instance, in a previous analysis that includes three *Schizosaccharomyces* species, *Schizosaccharomyces* plus Saccharomycotina group with strong support (BP 95%), although an alternative (likely correct) position of *Schizosaccharomyces* at the base of the Ascomycota was not rejected by an AU test (Bullerwell, Forget, and Lang 2003a). In this study, amino acid instead of nucleotide sequences have been used to decrease the effect of A+T bias, and we find no significant amino acid compositional bias in this data (result not shown). Yet, after inclusion of further complete mitochondrial data from three slowly-evolving Taphrinomycotina (*S. complicata*, *T. deformans*, and *P. carinii)*, the position of *Schizosaccharomyces* does not change, although the bootstrap support for this topology decreases to 76% (Fig 2). This result is consistent with the suggestion that adding more sequences (particularly from slowly-evolving species) usually helps to reduce the effect of LBA (for a review, see (Delsuc, Brinkmann, and Philippe 2005)).

We have further tested whether *Schizosaccharomyces* mitochondrial sequences contain little phylogenetic signal and a strong tendency for LBA by inferring a phylogeny with a distant fungal outgroup composed of four fast-evolving Chytridiomycota. In this case, *Schizosaccharomyces* changes its position, away from Saccharomycotina towards the base of Fungi, apparently due to LBA with the distantly related Chytridiomycota. When Saccharomycotina are removed from the original dataset, Taphrinomycotina become grouped together, though without significant support. Finally, positional sorting with the SF method confirms our interpretation. Only the slowest-evolving positions (S2-S5 data matrix) are able to recover the tree topology inferred with the nuclear dataset, although only with marginal statistical support. Our analyses strongly suggest that the grouping of *Schizosaccharomyces* with Saccharomycotina in trees based on mitochondrial data is due to a LBA artifact.

Limitations of mitochondrial sequence data in phylogenetic analysis

A limitation of mitochondrial genome data is their small data size compared to nuclear genomes. The most popular mitochondrial data set contains only 13 proteins, including some that are rather small (*atp9*, *nad4L*) and others that are fast-evolving (*nad2*, *nad6*) and are therefore of limited value for the inference of deep phylogenies. To expand these datasets, mitochondrial genes that were transferred to the nucleus might be added. Yet, because the A+T content and other evolutionary constraints are different in mitochondrial and nuclear genomes, evolutionary models and inference methods might have to be adapted.

Conclusion

The current analysis ends a long-standing controversy on the phylogenetic position of *Schizosaccharomyces* species: we conclude that they are part of Taphrinomycotina, branching at the base of Ascomycota. Yet, the phylogenetic identity of *Neolecta*, another putative representative of this group, remains to be assessed by phylogenomic analysis.

Supplementary Material

Figures S1-S3 and Table S1 are available at *Molecular Biology and Evolution* online

Acknowledgments

Salary and interaction support from the Canadian Institute for Advanced Research (BFL), the Canadian Institute of Health Research (BFL), the 'Bourses d'Excellence biT' (CIHR; YL) is acknowledged, and the 'Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación' (Government of Basque Country; NRE). JWL is supported by a Student Research Award from the Nova Scotia

Health Research Foundation. We thank Lise Forget and Tom Sesterhenn for help in mtDNA sequencing and sequence interpretation.

References

Anderson, J. S. 2001. The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the lepospondyli (Vertebrata, Tetrapoda). Syst Biol **50**:170-193.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972-977.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol **16**:817-825.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res **31**:1614-1623.

Burger, G., D. V. Lavrov, L. Forget, and B. F. Lang. 2007. Sequencing complete mitochondrial and plastid genomes. Nat Protoc **2**:603-614.

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

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

Delsuc, F., H. Brinkmann, H. Philippe, E. J. Douzery, E. A. Snell, E. Bapteste, O. Jeffroy, Y. Zhou, and N. Rodrigue. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet **6**:361-375.

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

Dutilh, B. E., V. van Noort, R. T. van der Heijden, T. Boekhout, B. Snel, and M. A. Huynen. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics **23**:815-824.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

Edman, J. C., J. A. Kovacs, H. Masur, D. V. Santi, H. J. Elwood, and M. L. Sogin. 1988. Ribosomal RNA sequence shows *Pneumocystis carinii* to be a member of the fungi. Nature **334**:519-522.

Eriksson, O. E., and K. Winka. 1997. Supraordinal taxa of the Ascomycota. Myconet **1**:1-16.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol **52**:665-673.

Fitzpatrick, D. A., M. E. Logue, J. E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol **6**:99.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52**:696-704.

Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G. Thorn, L. Tibell, W.

A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res **111**:509-547.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature **443**:818-822.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol **4**:18.

Koski, L. B., M. W. Gray, B. F. Lang, and G. Burger. 2005. AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics **6**:151.

Lang, B. F., and G. Burger. 2007. Purification of mitochondrial and plastid DNA. Nat Protoc **2**:652-660.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr Biol **12**:1773-1778.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol **7 Suppl 1**:S4.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095-1109.

Leigh, J., E. Seif, N. Rodriguez, Y. Jacob, and B. F. Lang. 2003. Fungal evolution meets fungal genomics. Pp. 145-161 *in* D. K. Arora, ed. Handbook of Fungal Biotechnology. Marcel Dekker Inc., New York.

Liu, Y. J., M. C. Hodson, and B. D. Hall. 2006. Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. BMC Evol Biol **6**:74.

Mar, J. C., T. J. Harlow, and M. A. Ragan. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol Biol **5**:8.

Nishida, H., and J. Sugiyama. 1993. Phylogenetic relationships among *Taphrina, Saitoella*, and other higher fungi. Mol Biol Evol **10**:431-436.

O'Brien, E. A., L. B. Koski, Y. Zhang, L. Yang, E. Wang, M. W. Gray, G. Burger, and B. F. Lang. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res **35**:D445-451.

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

Pramateftaki, P. V., V. N. Kouvelis, P. Lanaridis, and M. A. Typas. 2006. The mitochondrial genome of the wine yeast Hanseniaspora uvarum: a unique genome organization among yeast/fungal counterparts. FEMS Yeast Res **6**:77-90.

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

Rodriguez-Ezpeleta, N., H. Brinkmann, G. Burger, A. J. Roger, M. W. Gray, H. Philippe, and B. F. Lang. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol **17**:1420-1425.

Rodriguez-Ezpeleta, N., S. Teijeiro, L. Forget, G. Burger, and B. F. Lang. in press. Generation of cDNA libraries: protists and fungi *in* P. J., ed. Methods in Molecular Biology: Expressed Sequence Tags. Humana press Inc., Totowa, NJ.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol **7 Suppl 1**:S2.

Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol Biol Evol **20**:1036-1042.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol **51**:492-508.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**:1246-1247.

Spatafora, J. W., G. H. Sung, D. Johnson, C. Hesse, B. O'Rourke, M. Serdani, R. Spotts, F. Lutzoni, V. Hofstetter, J. Miadlikowska, V. Reeb, C. Gueidan, E. Fraker, T. Lumbsch, R. Lucking, I. Schmitt, K. Hosaka, A. Aptroot, C. Roux, A. N. Miller, D. M. Geiser, J. Hafellner, G. Hestmark, A. E. Arnold, B. Budel, A. Rauhut, D. Hewitt, W. A. Untereiner, M. S. Cole, C. Scheidegger, M. Schultz, H. Sipman, and C. L. Schoch. 2006. A five-gene phylogeny of Pezizomycotina. Mycologia **98**:1018-1028.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Sugiyama, J., K. Hosaka, and S. O. Suh. 2006. Early diverging Ascomycota: phylogenetic divergence and related evolutionary enigmas. Mycologia **98**:996-1005.

Swofford, D. L. 2000. PAUP*: phylogenetic analysis using parsimony and other methods. Sinauer, Sunderland, MA.

Taylor, D. J., and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol **21**:1534-1537.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol **52**:528-538.

Tables

Table 1: Likelihood tests of alternative tree topologies

945 topologies were generated by constraining well-supported internal branches (monophyly of Saccharomycotina, Pezizomycotina, Ascomycota, Basidiomycota, and Zygomycota plus Chytridiomycota as outgroup), leaving the four Taphrinomycotina unconstrained within Ascomycota. Table 1 lists the p-values of the 10 top-ranking topologies based on the AU test (data model as in Fig 1). The following abbreviations are used: *P.c: Pneumocystis carinii*; *S.c: Saitoella complicata*; *S.p: Schizosaccharomyces pombe*; *T.d: Taphrina deformans*; Sacch: Saccharomycotina; Pezi: Pezizomycotina. In the five best topologies, Taphrinomycotina are monophyletic. All other topologies in which they are paraphyletic are rejected at a significance level less than 0.01.

Rank	Tree topology	Taphrinomycotina	ΔlnL[a]	AU[b]
1	best tree (Figure 1)	Monophyletic	-14.4	0.869
2	((*T.d*, *S.c)*, (*S.p*, *P.c*)) at base of Ascomycota	Monophyletic	14.4	0.297
3	(*S.c*, (*P.c,* (*T.d, S.p*))) at base of Ascomycota	Monophyletic	27.6	0.131
4	((*T.d, S.p*), (*P.c*, *S.c*)) at base of Ascomycota	Monophyletic	45.1	0.032
5	(*P.c*, (*S.c*, (*S.p*, *T.d)*)) at base of Ascomycota	Monophyletic	50.2	0.011
6	(*T.d*, (*S.c*, (*S.p*, *P.c*)),(Sacch,Pezi))	Paraphyletic	163.1	**0.007**
7	((*S.p*, (*T.d*, *P.c))*, ((*S.c*, Sacch), Pezi))	Paraphyletic	525.6	**0.007**
8	((*T.d*, *S.c*), (*P.c*, (*S.p*, (Sacch, Pezi))))	Paraphyletic	243.0	**0.005**
9	(*S.p*, ((*S.c*, *T.d*), (*P.c*, (Sacch, Pezi))))	Paraphyletic	265.6	**0.004**
10	((*S.p*, *P.c*), ((*S.c, T.d*), (Sacch, Pezi)))	Paraphyletic	99.2	**0.004**

[a] Log likelihood difference

[b] Approximate Unbiased (AU) test

Figure legends

Figure 1: Phylogeny based on nucleus-encoded protein sequences

This tree was inferred from 113 nucleus-encoded proteins (29 387 amino acid positions), with three ML (Treefinder, PhyML and RAxML) and two BI (MrBayes and PhyloBayes) methods, either using the WAG+Gamma (four categories) model, or the CAT model (PhyloBayes). The PP using MrBayes and PhyloBayes are 1.0 for all branches, except for the one that groups *Taphrina* and *Saitoella* (PP 0.6). Numbers at internal branches represent support values obtained with 100 bootstrap replicates on the concatenated dataset with Treefinder/RAxML/PhyML. When all support values are identical, only one is indicated.

Figure 2: Phylogeny based on concatenated proteins encoded by mtDNA

The sequences of 13 proteins (*cox1, cox2, cox3, cob, atp6, atp9,* and *nad1, nad2, nad3, nad4, nad4L, nad5, nad6*) were concatenated (2 596 amino acid positions). See Figure 1 for details on inference methods. The BI with MrBayes has posterior probabilities of at least 0.99, except for the internal branch which groups *Allomyces* and other chytrids (PP 0.64).

Figure 3: Phylogenetic analysis of mitochondrial dataset after removing Saccharomycotina

All Saccharomycotina were removed from the complete mitochondrial dataset. The analyses were performed as in Figure 1. The *Schizosaccharomyces* group with other

Taphrinomycotina with various BPs among three ML methods (TreeFinder: 95%, RAxML: 66%, PhyML: 97%), the PP of BI using MrBayes is 1.0.

Figure 4: Impact of fast-evolving positions on the inferred phylogeny from proteins encoded by mtDNA.

The SF method was used to generate a series of 14 datasets (S0, S1, S2, …, S14) with an increasing fraction of fast-evolving sequence positions. The phylogenies were inferred using RAxML on these datasets (WAG+Gamma with four categories). Results with S3 and S5 are shown in Fig. 5A, and with S7 and S9 in Fig. 5B. Numbers at internal branches represent BP obtained with 100 bootstrap replicates, which are in the order S3, S5 (Fig. 5A) and S7, S9 (Fig. 5B) from top. When all bootstrap values are > 95%, only one value is presented.

Figure 1

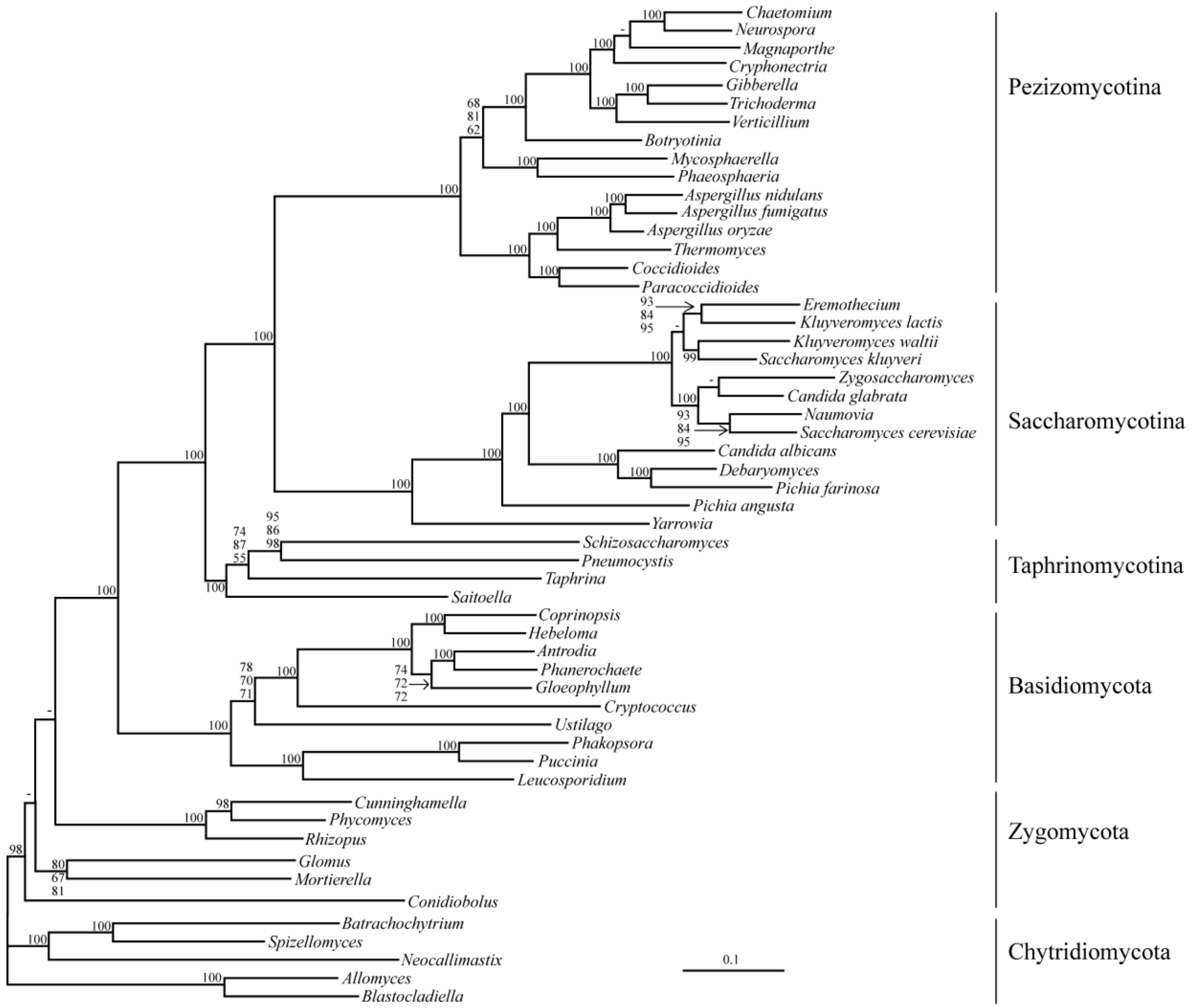

Chaetomium
Neurospora
Magnaporthe
Cryphonectria
Gibberella
Trichoderma
Verticillium
Botryotinia
Mycosphaerella
Phaeosphaeria
Aspergillus nidulans
Aspergillus fumigatus
Aspergillus oryzae
Thermomyces
Coccidioides
Paracoccidioides
Eremothecium
Kluyveromyces lactis
Kluyveromyces waltii
Saccharomyces kluyveri
Zygosaccharomyces
Candida glabrata
Naumovia
Saccharomyces cerevisiae
Candida albicans
Debaryomyces
Pichia farinosa
Pichia angusta
Yarrowia
Schizosaccharomyces
Pneumocystis
Taphrina
Saitoella
Coprinopsis
Hebeloma
Antrodia
Phanerochaete
Gloeophyllum
Cryptococcus
Ustilago
Phakopsora
Puccinia
Leucosporidium
Cunninghamella
Phycomyces
Rhizopus
Glomus
Mortierella
Conidiobolus
Batrachochytrium
Spizellomyces
Neocallimastix
Allomyces
Blastocladiella
0.1
Pezizomycotina
Saccharomycotina
Taphrinomycotina
Basidiomycota
Zygomycota
Chytridiomycota
68
81
62
93
84
95
93
84
95
95
86
98
74
87
55
78
70
71
74
72
72
80
67
81
98
99
98
-

Figure 2

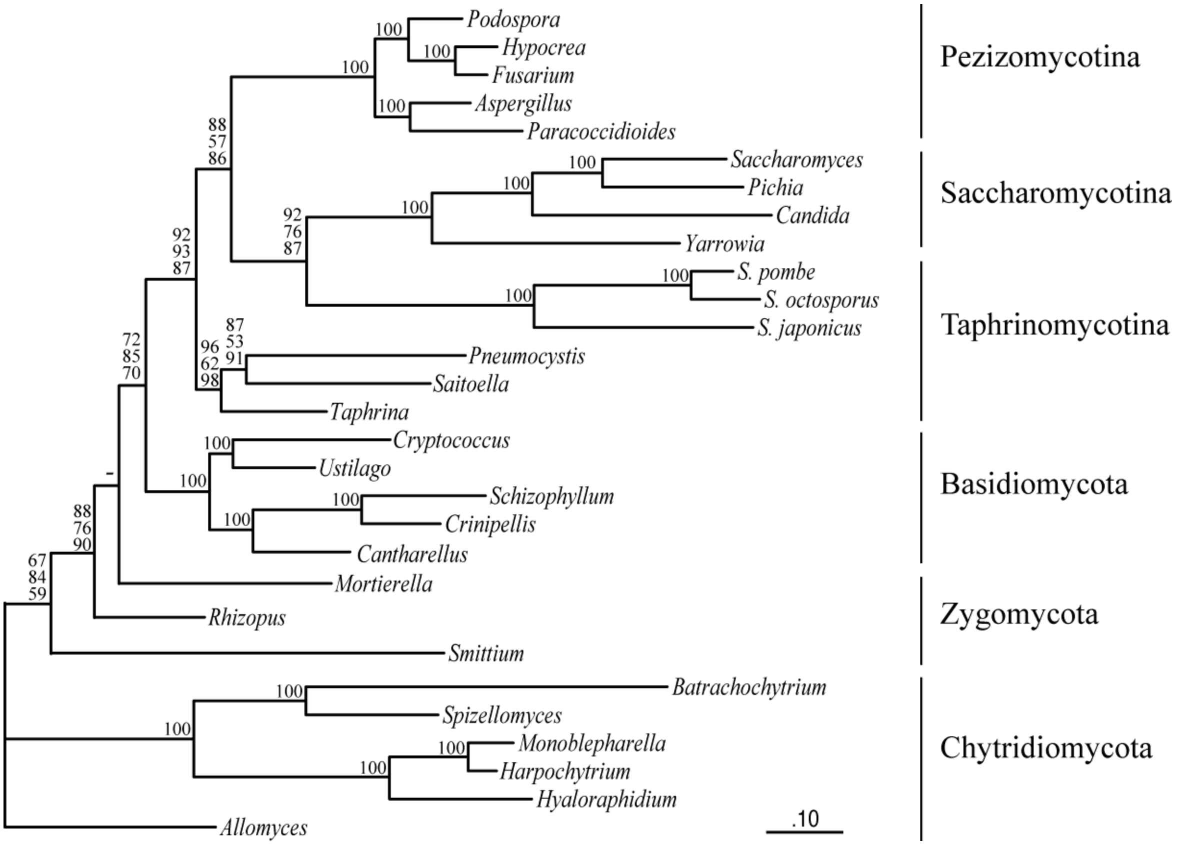

Podospora
Hypocrea
Fusarium
Aspergillus
Paracoccidioides
Saccharomyces
Pichia
Candida
Yarrowia
S. pombe
S. octosporus
S. japonicus
Pneumocystis
Saitoella
Taphrina
Cryptococcus
Ustilago
Schizophyllum
Crinipellis
Cantharellus
Mortierella
Rhizopus
Smittium
Batrachochytrium
Spizellomyces
Monoblepharella
Harpochytrium
Hyaloraphidium
Allomyces
100
100
100
100
88 57 86
100
100
100
92 76 87
92 93 87
100
100
87 53 91
96 62 98
72 85 70
100
100
100
100
-
88 76 90
67 84 59
100
100
100
100
.10
Pezizomycotina
Saccharomycotina
Taphrinomycotina
Basidiomycota
Zygomycota
Chytridiomycota

Figure 3

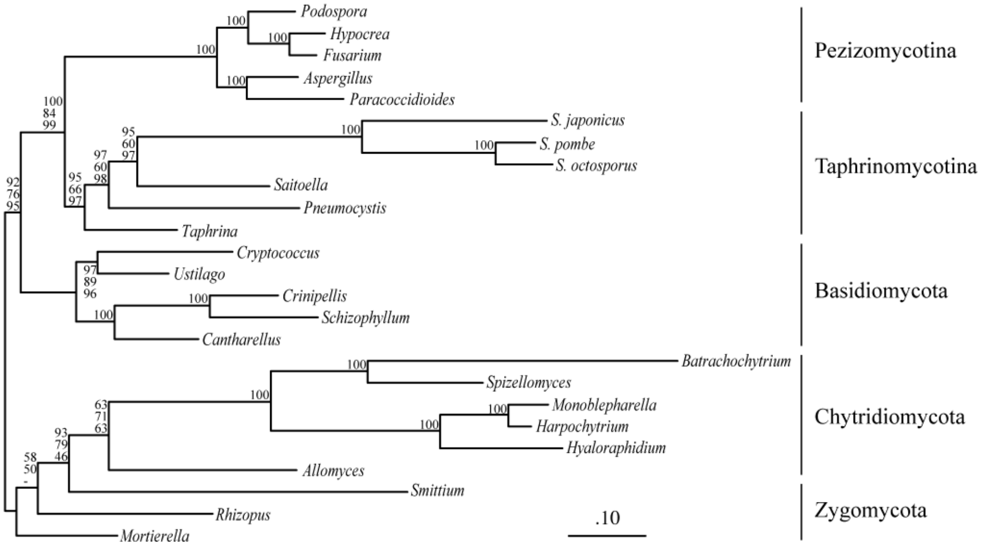
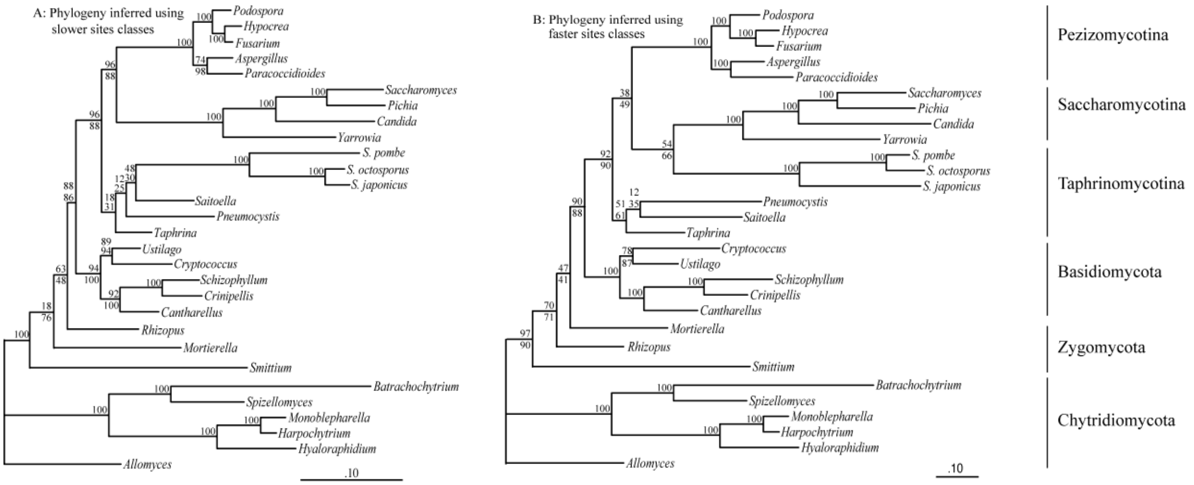

Podospora
100
100
Hypocrea
100
Fusarium
Aspergillus
100
Paracoccidioides
Pezizomycotina
100
84
99
S. japonicus
100
95
60
97
100
S. pombe
S. octosporus
97
60
98
Saitoella
95
66
97
Pneumocystis
92
76
95
Taphrina
Taphrinomycotina
Cryptococcus
97
89
96
Ustilago
100
Crinipellis
100
Schizophyllum
Cantharellus
Basidiomycota
100
Batrachochytrium
Spizellomyces
100
100
Monoblepharella
63
71
63
100
Harpochytrium
Hyaloraphidium
93
79
46
Allomyces
58
50
-
Smittium
Rhizopus
.10
Zygomycota
Mortierella
Chytridiomycota

Figure 4


A: Phylogeny inferred using slower sites classes
Podospora
Hypocrea
Fusarium
Aspergillus
Paracoccidioides
Saccharomyces
Pichia
Candida
Yarrowia
S. pombe
S. octosporus
S. japonicus
Saitoella
Pneumocystis
Taphrina
Ustilago
Cryptococcus
Schizophyllum
Crinipellis
Cantharellus
Rhizopus
Mortierella
Smittium
Batrachochytrium
Spizellomyces
Monoblepharella
Harpochytrium
Hyaloraphidium
Allomyces
.10
B: Phylogeny inferred using faster sites classes
Podospora
Hypocrea
Fusarium
Aspergillus
Paracoccidioides
Saccharomyces
Pichia
Candida
Yarrowia
S. pombe
S. octosporus
S. japonicus
Pneumocystis
Saitoella
Taphrina
Cryptococcus
Ustilago
Schizophyllum
Crinipellis
Cantharellus
Mortierella
Rhizopus
Smittium
Batrachochytrium
Spizellomyces
Monoblepharella
Harpochytrium
Hyaloraphidium
Allomyces
.10
Pezizomycotina
Saccharomycotina
Taphrinomycotina
Basidiomycota
Zygomycota
Chytridiomycota

SUPPLEMENTAL MATERIAL

Table S1: Proportion of missing positions in nuclear data set

The inclusion of an amount of missing positions allowed us to use more gene and species.

The proportion of missing positions for each species is listed in Table S1.

Species	Missing position(%)	Species	Missing position(%)
Allomyces macrogynus	0.504	Mycosphaerella graminicola	0.503
Antrodia cinnamomea	0.296	Naumovia castellii	0.117
Aspergillus fumigatus	0.049	Neocallimastix patriciarum	0.611
Aspergillus nidulans:	0.018	Neurospora crassa	0
Aspergillus oryzae:	0.003	Paracoccidioides brasiliensis	0.307
Batrachochytrium dendrobatidis	0.048	Phaeosphaeria nodorum	0.081
Blastocladiella emersonii	0.207	Phakopsora pachyrhizi	0.279
Botryotinia fuckeliana	0.348	Phanerochaete chrysosporium	0.105
Candida albicans	0.009	Phycomyces blakesleeanus	0.029
Candida glabrata	0.017	Pichia angusta	0.619
Chaetomium globosum	0.095	Pichia farinose	0.760
Coccidioides immitis	0.012	Pneumocystis carinii	0.380
Conidiobolus coronatus	0.810	Puccinia graminis	0.013
Coprinopsis cinerea	0.008	Rhizopus oryzae	0.006
Cryphonectria parasitica	0.517	Saccharomyces cerevisiae	0.003
Cryptococcus neoformans	0.016	Saccharomyces kluyveri	0.177
Cunninghamella elegans	0.346	Saitoella complicata	0.667
Debaryomyces hansenii	0.036	Schizosaccharomyces pombe	0.014
Eremothecium gossypii	0.014	Spizellomyces punctatus	0.434
Gibberella zeae	0.011	Taphrina deformans	0.568
Gloeophyllum trabeum	0.392	Thermomyces lanuginosus	0.412
Glomus intraradices	0.694	Trichoderma reesei	0.311
Hebeloma cylindrosporum	0.491	Ustilago maydis	0.023
Kluyveromyces lactis	0.039	Verticillium dahliae	0.614
Kluyveromyces waltii	0	Yarrowia lipolytica	0.015
Leucosporidium scottii	0.373	Zygosaccharomyces rouxii	0.709
Magnaporthe grisea	0.027	Mortierella verticillata	0.562

Figure S1: Phylogeny based on nucleus-encoded proteins with less missing data

The analysis is based on 76 nucleus-encoded proteins, with a reduced proportion of missing positions for *S. complicata* (43.0%) and *T. deformans* (39.9%). This tree was inferred with ML (Treefinder, RAxML and PhyML) and BI (MrBayes and PhyloBayes) methods either using the WAG+Gamma (four categories) model or the CAT model (PhyloBayes). The PP of BI analyses are 1.0 for all branches, except for the internal branch that groups *Taphrina*, *Schizosaccharomyces* and *Pneumocystis* (PP: 0.73) and the branch that groups *Mortierella* and *Conidiobollus* (PP: 0.75). Numbers around the internal nodes represent support values obtained with 100 bootstrap replicates of the concatenated dataset with Treefinder/RAxML/PhyML (WAG+Gamma model). When all support values were identical, only one is indicated.

Figure S2: Phylogenetic analysis of mitochondrial dataset containing only Ascomycota and four fast-evolving Chytridiomycota

The mitochondrial dataset was reduced to 19 species, including all 15 Ascomycota plus four (fast-evolving) Chytridiomycota. See Figure 1 for the details of inference methods and legend. Note that instead of grouping with Saccharomycotina, the *Schizosaccharomyces* were attracted to the base of Fungi by fast-evolving Chytridiomycota. The BI with MrBayes has posterior probabilities of 1.0, except for the internal branch that groups *Pneumocystis*, *Saitoella*, *Taphrina*, Saccharomycotina and Pezizomycotina(PP: 0.5) and the branch that groups *Schizosaccharomyces* with all other Ascomycota (PP: 0.95).

Figure S3: Bootstrap supports for two alternative positions of *Schizosaccharomyces* (one is grouping with other Taphrinomycotina, the other grouping with Saccharomycotina) using SF method

Different subsets of the mitochondrial dataset (S2, S3, …, S14) were constructed (see Methods and Material for detail). ML analyses were performed using RAxML. The most slowly evolving subsets (S2-S5) did not significantly support either topology, although the topology with *Schizosaccharomyces* grouping with other Taphrinomycotina (the topology supported by nuclear data) has better likelihood value. As more fast-evolving positions were included (the S6 - S14 datasets), the other topology with *Schizosaccharomyces* grouping with **Saccharomycotina** (the incorrect one) was recovered with a higher BV.

Figure S1

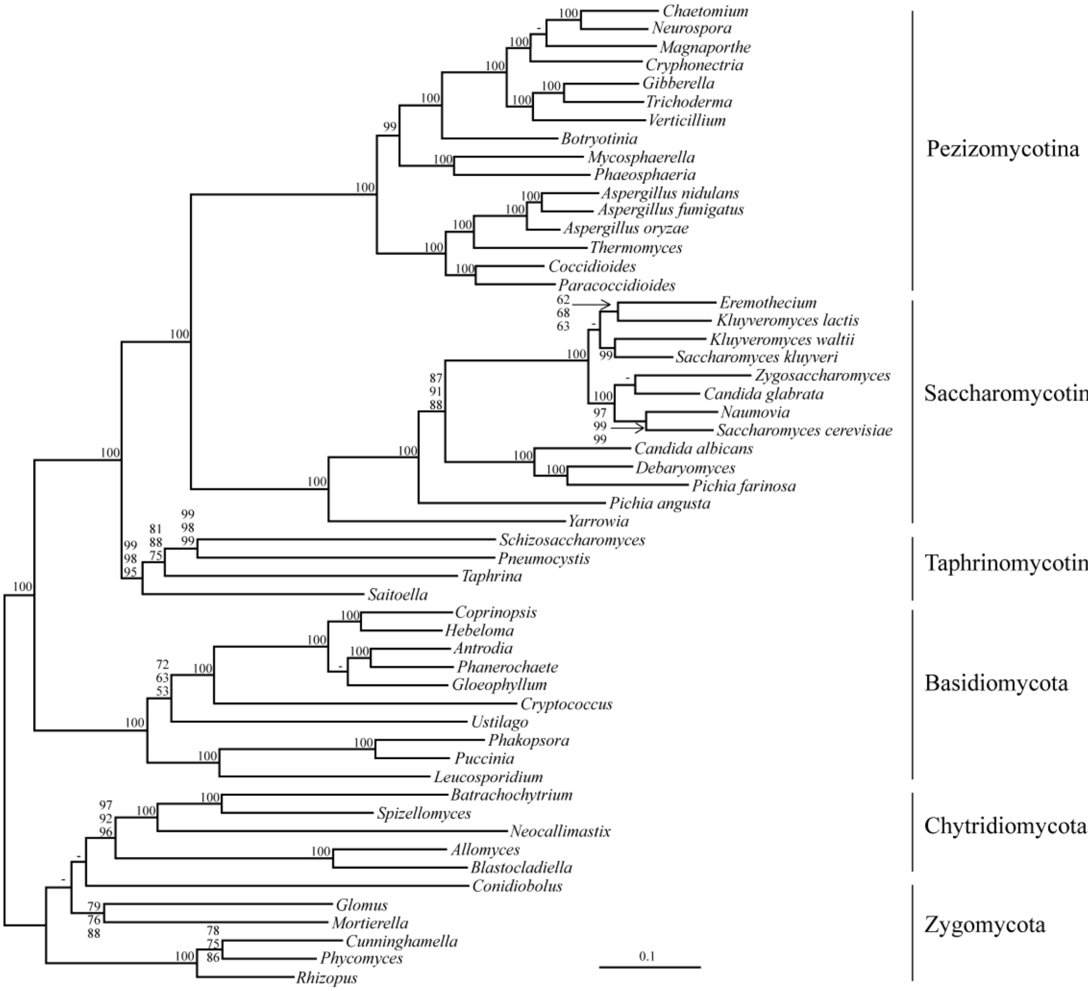

100
Chaetomium
-
Neurospora
100
Magnaporthe
100
Cryphonectria
100
Gibberella
100
Trichoderma
Verticillium
100
99
Botryotinia
100
Mycosphaerella
Phaeosphaeria
100
100
Aspergillus nidulans
100
Aspergillus fumigatus
100
Aspergillus oryzae
100
Thermomyces
100
Coccidioides
Paracoccidioides
Pezizomycotina
62
68
63
Eremothecium
-
Kluyveromyces lactis
Kluyveromyces waltii
100
99
Saccharomyces kluyveri
87
91
88
-
Zygosaccharomyces
Candida glabrata
100
97
99
99
Naumovia
Saccharomyces cerevisiae
100
Candida albicans
100
100
Debaryomyces
100
Pichia farinosa
Pichia angusta
Yarrowia
Saccharomycotina
100
100
99
98
99
81
88
75
99
98
95
Schizosaccharomyces
Pneumocystis
Taphrina
Saitoella
Taphrinomycotina
100
100
Coprinopsis
Hebeloma
100
100
Antrodia
-
Phanerochaete
Gloeophyllum
72
63
53
100
Cryptococcus
Ustilago
100
100
Phakopsora
Puccinia
100
Leucosporidium
Basidiomycota
97
92
96
100
100
Batrachochytrium
Spizellomyces
Neocallimastix
100
Allomyces
Blastocladiella
-
-
Conidiobolus
Chytridiomycota
79
76
88
Glomus
Mortierella
78
75
86
Cunninghamella
100
Phycomyces
Rhizopus
0.1
Zygomycota

Figure S2

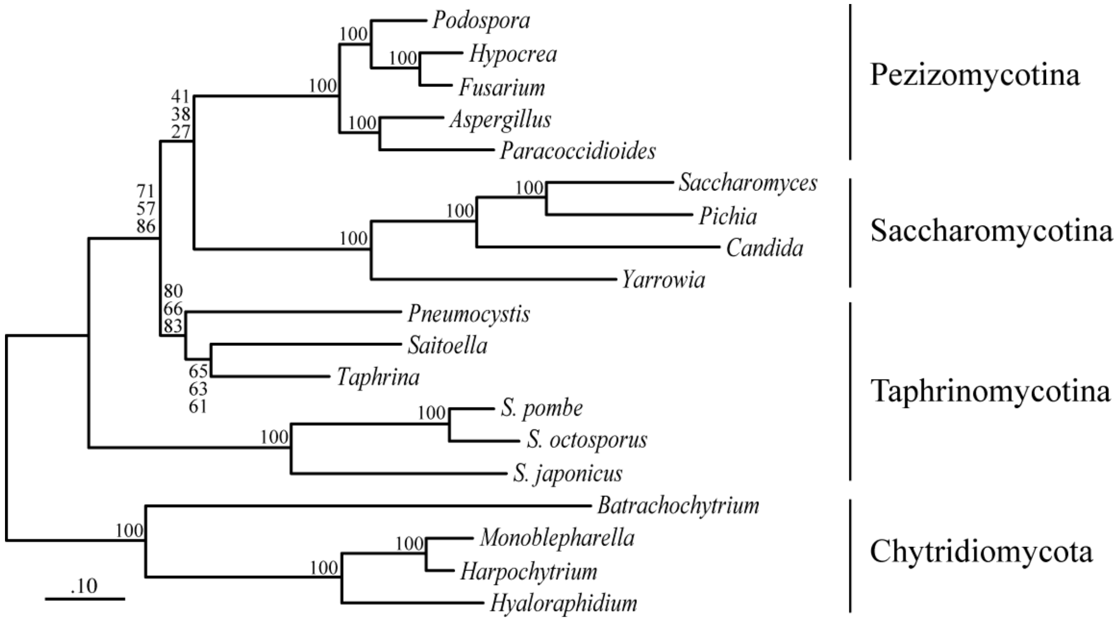
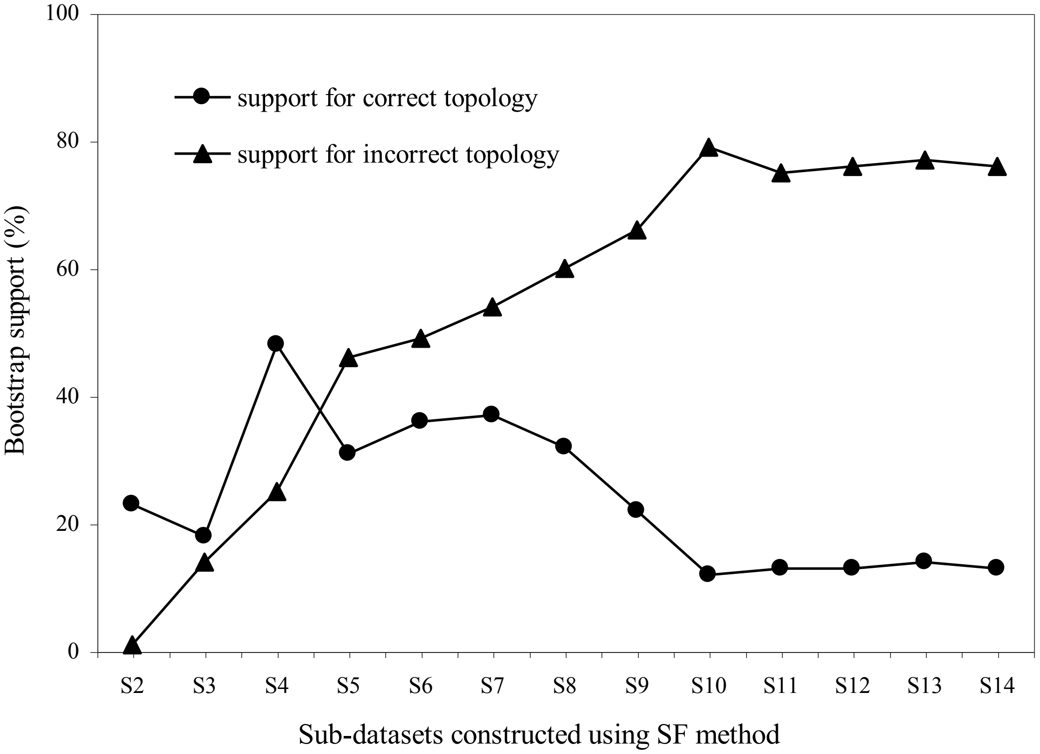

Podospora
100
100
Hypocrea
Fusarium
100
41
38
27
100
Aspergillus
Paracoccidioides
Pezizomycotina
Saccharomyces
100
Pichia
100
71
57
86
100
Candida
Yarrowia
Saccharomycotina
80
66
83
Pneumocystis
Saitoella
65
63
61
Taphrina
S. pombe
100
100
S. octosporus
S. japonicus
Taphrinomycotina
Batrachochytrium
100
100
Monoblepharella
100
Harpochytrium
Hyaloraphidium
Chytridiomycota
.10

Figure S3


100
80
60
40
20
0
Bootstrap support (%)
support for correct topology
support for incorrect topology
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
Sub-datasets constructed using SF method

Chapter 4 Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support

Yu Liu[1,3]*, Emma T. Steenkamp[2]*, Henner Brinkmann[1], Lise Forget[1], Hervé Philippe[1] and B. Franz Lang[1]

[1]Robert Cedergren Centre, Département de biochimie, Université de Montréal, Montréal, Québec, Canada.

[2]Department of Microbiology and Plant Pathology, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa.

[3]Present address: *Donnelly Centre for Cellular and Bio-molecular Research, Department of Molecular Genetics, University of Toronto, 160 College Street, Toronto, ON, Canada, M5S 3E1*

* These authors contributed equally to this paper

Corresponding author: B. Franz Lang

SUMMARY

Background: Resolving the evolutionary relationships among Fungi remains challenging because of their highly variable evolutionary rates, and lack of a close phylogenetic outgroup. Nucleariida, an enigmatic group of amoeboids, have been proposed to emerge close to the fungal-metazoan divergence and might fulfill this role. Yet, published phylogenies with up to five genes are without compelling statistical support, and genome-level data should be used to resolve this question with confidence.

Results: Our analyses with nuclear (118 proteins) and mitochondrial (13 proteins) data now robustly associate Nucleariida and Fungi as neighbors, an assemblage that we term 'Holomycota'. With Nucleariida as an outgroup, we revisit unresolved deep fungal relationships.

Conclusions: Our phylogenomic analysis provides significant support for the paraphyly of the traditional taxon Zygomycota, and contradicts a recent proposal to include *Mortierella* in a phylum Mucoromycotina. We further question the introduction of separate phyla for Glomeromycota and Blastocladiomycota, whose phylogenetic positions relative to other phyla remain unresolved even with genome-level datasets. Our results motivate broad sampling of additional genome sequences from these phyla.

BACKGROUND

The investigation of previously little known eukaryotic lineages within and close to the opisthokonts will be key to understanding the origins of Fungi, the evolution of developmental traits in Fungi and Metazoa, and ultimately the origin(s) of multicellularity (Kaiser 2001; Keeling et al. 2005; Ruiz-Trillo et al. 2007). In particular, it will help to establish which and how many developmental genes are either shared or specific to these two major eukaryotic groups. In this context, it is essential to determine the precise phylogenetic position of candidate protists that are close to Fungi, Metazoa, or opisthokonts as a whole.

The candidate organisms choanoflagellates, ichthyosporeans and *Ministeria* have been convincingly shown to be relatives of Metazoa (combined in a taxon termed Holozoa; (Lang et al. 2002b)) by using molecular phylogenetics with genomic datasets (e.g., (Lang et al. 2002b; Ruiz-Trillo et al. 2006; Jimenez-Guri et al. 2007; Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008). Yet, there are remaining questions about the exact phylogenetic positions of *Capsaspora* (Jimenez-Guri et al. 2007; Ruiz-Trillo et al. 2008) and *Ministeria* (Shalchian-Tabrizi et al. 2008) within Holozoa. Another, less well studied group of protists are Nucleariida, a group of heterotrophic amoeboids with radiating filopodia. Nucleariids lack distinctive morphological features that might allow associating them with either animals or fungi. Their mitochondrial cristae are either discoidal-shaped or flattened (Patterson 1984; Patterson 1999; Amaral-Zettler et al. 2001). Indeed, initial phylogenetic analyses based on single genes have been inconsistent in placing them even within opisthokonts. There has been also confusion due to the inclusion within Nucleariida of *Capsaspora owczarzaki*, a species that is now excluded from this group and shown to be clearly associated with Holozoa (Amaral-Zettler et al. 2001; Hertel, Bayne, and Loker 2002; Cavalier-Smith and Chao 2003a; Dykova, Fiala, and Peckov 2003; Medina 2003; Dykova and Lom 2004; Nikolaev et al. 2004; Ruiz-Trillo et al. 2008).

Overall, the phylogenetic position of the ‘true’ nucleariids remains controversial. In a more recent phylogenetic investigation with four nuclear gene sequences (EF-1α, HSP70, actin and β-tubulin), nucleariids associate confidently with Fungi, but only when selecting two slow-evolving chytridiomycetes (Steenkamp, Wright, and Baldauf 2006). When improving the taxon sampling to 18 fungal species, the bootstrap support (BS) value for fungal monophyly drops to 85 %, and alternative nucleariid positions are not rejected with the approximately unbiased (AU) test (Shimodaira 2002; Steenkamp, Wright, and Baldauf 2006). In this context, it seems noteworthy that *Nuclearia* and fungi other than chytrids are fast-evolving, and that the rate of tubulin evolution varies strongly among species of the latter dataset (correlating to some degree with the independent loss of the flagellar apparatus in non-chytrid fungi and in *Nuclearia*). Together, these rate differences at the gene and species levels may increase long-branch-attraction (LBA between the two fast-evolving groups) thus causing weaker support for fungal monophyly and the nucleariid-fungal sister relationship, or predicting altogether incorrect phylogenetic relationships.

These unresolved questions served as motivation for the current phylogenetic analyses that are based on broad taxon sampling, substantially more nuclear genes (available through expressed sequence tag (EST) or complete genome projects), and comparative analyses of nuclear and mitochondrial gene datasets. To this end, we sequenced several thousand ESTs each from two *Nuclearia simplex* strains (representing most distant members of this group rather than the same species, based on the high level of sequence divergence between them), and added them to a previous dataset (Rodriguez-Ezpeleta et al. 2007a) along with new genome data available from Holozoa (*C. owczarzaki, Amoebidium parasiticum, Sphaeroforma arctica*; (Ruiz-Trillo et al. 2008)) and Fungi (*Allomyces macrogynus, Batrachochytrium dendrobatidis,* and *Mortierella verticillata*). We then sequenced the mitochondrial genome of one of the two *N. simplex* strains. Similar to the nuclear genomes of fungi, their mitochondrial genomes also evolve at varying rates thereby introducing a considerable potential for phylogenetic artifacts. However, phylogenetic comparisons between mitochondrial and nuclear data provide valuable, cross-wise indicators of phylogenetic artifacts as the respective evolutionary rates differ between the two genomes. For instance, such comparisons have revealed inconsistencies for the positioning of *Schizosaccharomyces* species within Taphrinomycotina (Liu et al. 2009a), and of *Capsaspora* within Holozoa (Jimenez-Guri et al. 2007; Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008).

If the nucleariids are indeed the closest known relatives of Fungi as claimed (Steenkamp, Wright, and Baldauf 2006), this protist group will provide an excellent fungal outgroup that would ultimately facilitate the settling of controversial phylogenetic placement of taxa within Fungi and/or in close neighboring groups. Among the debated issues are the monophyly and appropriate classification of the traditional fungal taxa Chytridiomycota and Zygomycota. Previous analyses based on single or a few genes have been inconsistent in answering these questions, and often lack significant support (James et al. 2000; Tehler, Little, and Farris 2003; Taylor et al. 2004; Seif et al. 2005; Tanabe, Watanabe, and Sugiyama 2005b; James et al. 2006a; James et al. 2006b; Liu, Hodson, and Hall 2006; Spatafora et al. 2006; Hibbett et al. 2007). For example, the analyses of ribosomal RNA data supports the sister relationship between Glomeromycota and Dikarya (Ascomycota plus Basidiomycota) (Tehler, Little, and Farris 2003), while analysis of genes encoding the largest and second-largest subunits of the nuclear RNA polymerase II supports the monophyly of Zygomycota in its traditional definition (Liu, Hodson, and Hall 2006).

Phylogenetic positioning of the extremely fast-evolving Microsporidia (causing strong LBA attraction artifacts in phylogenetic analyses) is another controversial issue of great interest. In some of the more recent analyses, Microsporidia have been placed either close to zygomycetes/Mucorales (Keeling 2003; Lee et al. 2008), or together with *Rozella allomycis* (James et al. 2006a). Together with environmental sequences, *Rozella* species form part of a large, diverse and relatively slowly evolving lineage (designated "Rozellida"). They branch as a sister clade to Fungi (James et al. 2006a; Lara, Moreira, and Lopez-Garcia 2009b), which raises the additional question whether they should be considered to be true fungi as originally proposed (Adl et al. 2005). Testing the above alternative hypotheses on microsporidian affinities by phylogenomic analysis will require much more data from Rozellida (a few genes are known from *Rozella allomycis*, but largely insufficient for inclusion in our analyses), and from a much wider range of the paraphyletic zygomycetes. Generation of genome-size data will be further critical for applying methods that reduce LBA artifacts such as removal of fast-evolving genes or sequence sites (e.g., and refeences therein (Rodriguez-Ezpeleta et al. 2007b)).

Despite these and various other unresolved phylogenetic issues, fungal taxonomy has been substantially redefined in a recent proposal (Hibbett et al. 2007). Chytridiomycota is still treated as a phylum, but now include only Chytridiomycetes and Monoblepharidomycetes. Other traditional chytrid lineages such as Blastocladiomycota and Neocallimastigales have been elevated to phyla based on the analyses of LSU and SSU rRNA (James et al. 2006b), although support with these and other molecular markers is inconclusive. In turn, the traditional phylum Zygomycota has been altogether removed from this taxonomy (Hibbett et al. 2007), because evolutionary relationships among its members are currently unresolved and suspected to be paraphyletic. Zygomycota are now reassigned into a phylum Glomeromycota plus four subphyla *incertae sedis* (i.e., uncertain): Mucoromycotina, Kickxellomycotina, Zoopagomycotina and Entomophthoromycotina. To revisit these somewhat contentious issues, we compared results with mitochondrial and nuclear phylogenomic datasets, and further analyzed the effect of extending fungal species sampling, with the two *N. simplex* strains as the outgroup.

RESULTS AND DISCUSSION

Phylogenomic analysis with the eukaryotic dataset supports Nucleariida as sister to Fungi.

Phylogenomic analysis of the eukaryotic dataset with one of the currently most realistic phylogenetic models (category mixture model (CAT); (Lartillot and Philippe 2004)) confirms the monophyly of major eukaryotic groups including Holozoa, Fungi, Amoebozoa, and Viridiplantae. Further, *Amoebidium*, *Sphaeroforma* plus *Capsaspora* form a monophyletic group, and *Nuclearia* is without a doubt the closest known sister-group to Fungi (100% BS; Fig. 1). Also some higher-order relationships are recovered with significant support, such as opisthokonts and the two recently proposed supergroups JEH (jakobids, Euglenozoa plus Heterolobosea (Rodriguez-Ezpeleta et al. 2007a)) and CAS (Cercozoa, Alveolata plus Stramenopila (Hackett et al. 2007; Rodriguez-Ezpeleta et al. 2007a; Burki, Shalchian-Tabrizi, and Pawlowski 2008)), whereas monophyly of Plantae, Excavata and Chromalveolata is not found. Evidently, the taxon sampling of protists in our dataset is insufficient for (and not aimed at) resolving the phylogenetic relationships among these latter lineages, as it was meant to constitute only a strong and well sampled outgroup to opisthokonts.

Analysis of the eukaryotic dataset with maximum likelihood (ML) using RAxML (Stamatakis 2006) and the commonly used WAG+Γ model generated a similar tree topology (Fig. 1 and Fig. S1). Deep opisthokont divergences are predicted consistently and with significant support (BS > 98%), with *Nuclearia* clearly sister to Fungi (100% BS) and choanoflagellates the closest neighbor of animals. *Amoebidium*, *Sphaeroforma* plus *Capsaspora* form a monophyletic sister group to animals plus choanoflagellates, consistent with a previous analysis (Ruiz-Trillo et al. 2008) but contradicting others (Jimenez-Guri et al. 2007; Shalchian-Tabrizi et al. 2008). The reasons for this incongruence may be related to differences in data and taxon sampling. Our dataset contains 50 eukaryotic species with a close outgroup to Holozoa (i.e., including nucleariids together with fungal representatives), compared with a total of only 30 species in a more extensive previous analysis (Shalchian-Tabrizi et al. 2008). In contrast to our Bayesian analysis (BI), ML associates Malawimonadozoa with JEH (77% BS), a tendency noted and discussed previously (Rodriguez-Ezpeleta et al. 2007a; Hampl et al. 2009), and an issue to be addressed by better taxon sampling in this group (currently, data are available from only two species). Other minor differences between WAG *versus* CAT model analyses (yet without statistical support in favor of alternatives) are in relationships within Plantae and the placement of Haptophyceae.

We further investigated if the position of *Nuclearia* next to Fungi might be affected by potential phylogenetic artifacts, such as compositional sequence bias and/or LBA (Felsenstein 1978a; Rodriguez-Ezpeleta et al. 2007b). This is suspected because of the highly varying evolutionary rates both within Fungi and in protist outgroups, and the unusual result that better taxon sampling in Fungi reduces phylogenetic support for the *Nuclearia* position ((Steenkamp, Wright, and Baldauf 2006); see introduction). To do so, we first eliminated fast-evolving species from the dataset: *S. cerevisiae*, *Blastocystis* *hominis*, *Cryptosporidium parvum*, *Sterkiella histriomuscorum*, *Diplonema papillatum* and *Leishmania major*. The results from analyses using RAxML were essentially unchanged, both with respect to tree topology and BS values (Supplementary Material, Fig. S2). To counteract sequence bias, we recoded the 20 amino acids into six groups as previously proposed (Hrdy et al. 2004). Again, phylogenetic analysis of this dataset using P4 (Foster 2004) generated essentially the same tree topology, with some BS values slightly decreased due to the loss of information by recoding. (Supplementary Material, Fig. S3).

Finally, we evaluated the positioning of *Nuclearia* next to Fungi with the AU and weighted Shimodeira Hasegawa (wSH) likelihood tests (Shimodaira and Hasegawa 2001). For this, we compared the topology presented in Fig. 1 with competing tree topologies in which the two *Nuclearia* strains were moved as sistergroup to all major eukaryotic lineages, and all possible positions within Opisthokonta. The results of both tests confirm *Nuclearia* as the closest neighbor group of Fungi, with all alternative topologies rejected at a significance level of p=0.002 (Table 1). Given the unequivocal support for *Nuclearia* as the fungal sistergroup, we propose the term ‘Holomycota’ to refer to the assemblage of Nucleariida plus Fungi.

Mitochondrial phylogeny and genomic features support monophyly of the Holomycota.

Phylogenetic analyses of nuclear *versus* mitochondrial datasets are expected to come to similar conclusions, thus providing independent evidence for the given phylogenetic relationships. To this end, we sequenced and analyzed the complete mitochondrial DNA (mtDNA) of one of the *N. simplex* strains (a circular mapping DNA of 74 120 bp; see Supplementary Material, Fig. S4). Note that growth of *Nuclearia* is complicated (the standard method calls for growth on Petri dishes with a bacterial lawn as food source), and that it is difficult to obtain sufficient cell material for mtDNA purification, explaining why we succeeded for only one of the two *Nuclearia* species.

The *Nuclearia* mtDNA contains a high number of introns (21 group I, and one group II), and mitochondrial protein genes appear to be translated with the standard translation code. These features are also widespread in Fungi. In contrast, Holozoa all use a mitochondrial UGA (tryptophan) codon reassignment, and contain no or only a few introns (with the notable exception of Placozoa, an enigmatic group of Metazoa (Dellaporta et al. 2006)).

Phylogenetic analysis of a dataset with 56 species and 13 of the ubiquitous, most conserved mtDNA-encoded proteins predicts the monophyly of Opisthokonta, Stramenopila, Holozoa and Fungi with confidence, and also recovers *Nuclearia* as the sister-group of Fungi, albeit with a moderate BS value of 85% (Fig. 2). To verify if the limited support for Holomycota is expected (i.e., correlating with the number of available sequence positions in the respective datasets), we performed a variable length bootstrap (VLB) analysis. It compares the development of BS values with the number of sequence positions, for the nucleariid/fungal sister relationship. For this, we chose the 29 species shared between the two datasets (for the tree topology of the respective nuclear dataset see Supplementary Material, Fig. S5). The results show that the development of BS values is similar for nuclear and mitochondrial data (Fig. 3), and that the available mitochondrial dataset (as well as the above-cited nuclear phylogenies with five genes) is too small to resolve the phylogenetic position of nucleariids with high confidence. A better taxon sampling primarily in nucleariids will be imperative for improved phylogenetic resolution, motivating sequencing projects with new technologies, which are likely to provide mitochondrial and nuclear genome sequences - even with the limited amount of cellular material that is available for some taxa (e.g., (Lee and Young 2009)).

Fungal phylogeny with Nucleariida as outgroup.

Analyses of both the nuclear and mitochondrial datasets have been insufficient to assess with confidence, neither zygomycete mono/paraphyly, nor the phylogenetic position of Blastocladiomycota (Blastocladiales) (Fig. 1,2). For instance, a recent mitochondrial multi-gene phylogeny with the first complete *Glomus* mtDNA sequence groups *Glomus* and *Mortierella*, yet lacks significant statistical support (Lee and Young 2009). To re-address these questions, we have assembled a large dataset of nuclear-encoded genes from an extended, representative selection of fungal species, plus the two *Nuclearia* species as outgroup (i.e., the fungal dataset). The analyses show overall strong BS for the paraphyly of zygomycetes (Fig. 4), i.e. the Entomophthoromycotina represent a significantly supported and completely independent fungal lineage. However, monophyletic Mucoromycotina including *Mortierella* as recently redefined (Hibbett et al. 2007) is not recovered (rendering the taxon Mucoromycotina paraphyletic), neither is the taxon Symbiomycota (Glomeromycota plus Dikarya; (Tehler, Little, and Farris 2003)). Instead, there is moderate support to group Mucorales plus Dikarya (92% BS in BI) and *Glomus* as their next neighbor (85% BS in BI). Although the placement of *Glomus* relative to *Mortierella* differs between our BI and ML analyses (Fig. 4), we assume that the result of the BI analysis with its superior evolutionary model is more reliable. In light of these results, taxonomic reordering based on stable phylogenetic resolution of the traditional zygomycetes will require phylogenomic analyses with a much improved taxon sampling. Currently, nuclear and mitochondrial genome data are available only for single species in the latter two taxa; i.e. *Glomus intraradices* and *Mortierella verticillata*.

Rooting of the fungal tree with nucleariids confirms that the traditional chytridiomycetes are also paraphyletic, again assuming that the result of the BI analysis is correct (Fig. 4). Confirmation of this result (justifying an elevation of Blastocladiomycota as a separate phylum; (Hibbett et al. 2007)) is highly desirable, as genome-size datasets in Blastocladiomycota are limited to the two moderately distant species *Blastocladiella emersonii* and *Allomyces macrogynus*. Similarly, in light of the significant support for a monophyletic Chytridiomycota plus Neocallimastigomycota (100% BS with BI; Fig. 4), their division into separate taxonomic higher ranks should be reconsidered, but only after phylogenomic analysis with improved taxon sampling in both groups. Finally, our results motivate genome sequencing in *Rosella* species (Rozellida), potential relatives of Microsporidia and close neighbors of Fungi. The availability of a largely improved taxon sampling in zygomycetes, chytrids and Rozellida will provide a solid basis for evaluating the proposed placements of Microsporidia - either within or as a sistergroup to Fungi - based on phylogenomic analyses.

The results presented here are consistent with previous notions on how Fungi came into being. For example it is thought that the first Fungi probably had branched chytrid-like rhizoids, which developed by enclosure of nucleariid-like filopodia (sometimes branched) into cell walls, during a nutritional shift from phagotrophy to saprotrophy, thus giving rise to fungal hyphae and rhizoids (Shalchian-Tabrizi et al. 2008). However, the picture is more complicated as it is widely thought that the ancestral opisthokont also had a single posterior flagellum (Cavalier-Smith 1987a). This structure was lost during evolution of most but not all fungal lineages (e.g., (Patterson 1999; Berbee and Taylor 2000; Redecker 2002; Liu, Hodson, and Hall 2006), with a separate loss in the nucleariid sistergroup. In this sense, nucleariids are unlikely to represent a primitive developmental stage, but rather a secondary reduction resulting in a unicellular, amoeboid life style. Obviously, the clarification of the chain of events leading to the emergence of multicellularity in Fungi is by no means complete. These issues will only become clear with a much broader sampling of genomes from taxa near the animal-fungal divergence and the discovery of additional protist groups that are closely related to Fungi.

Conclusions

Here we demonstrate that phylogenomic analysis with improved evolutionary models and algorithms has a potential for resolving long-standing issues in fungal evolution, by increasing phylogenetic resolution. Yet, while our results support certain aspects of the new taxonomic classification of Fungi they contradict others, suggesting that the introduction of certain higher-level taxa is only preliminary. In particular, the elevation of Neocallimastigales, Blastocladiomycota and Glomeromycota to separate phyla is questionable from a molecular phylogenetics standpoint, and potentially confusing to the larger scientific community. At present, genome analyses continue to suffer from poor sampling in chytrids, zygomycetes and close fungal relatives such as nucleariids. This issue will be resolved by the employment of new, increasingly inexpensive genome sequencing technologies. Phylogenomic projects like the current one will help focusing on genome analyses of poorly known phyla and taxa that are key to understanding fungal origins and evolution.

Materials and Methods

Construction of cDNA libraries and EST sequencing

Two *N. simplex* (CCAP 1552/2 and 1552/4) cDNA libraries were constructed following recently published protocols (Rodriguez-Ezpeleta et al. 2009). Cells were grown in liquid standing cultures in WCL medium (http://megasun.bch.umontreal.ca/People/lang/FMGP/methods/wcl.html) supplemented with 0.5 x Cerophyll, with *E. coli* cells as food, which were pre-grown on LB medium in Petri-dishes as food. Plasmids were purified using the QIAprep 96 Turbo Miniprep Kit (Qiagen), sequencing reactions were performed with the ABI Prism BigDye^{TM} terminator version 3.0/3.1 (Perkin-Elmer, Wellesley, MA, USA) and sequenced on an MJ BaseStation (MJ Research, USA). Trace files were imported into the TBestDB database (http://tbestdb.bcm.umontreal.ca/searches/login.php) (O'Brien et al. 2007) for automated processing, including assembly as well as automated gene annotation by AutoFact (Koski et al. 2005b; Shen et al. 2009).

Mitochondrial sequencing and genome annotation

N. simplex (CCAP 1552/2) was grown as described above. The harvested cells were disrupted by addition of SDS plus proteinase K, and mitochondrial DNA was purified following a whole cell lysate protocol (Lang and Burger 2007) and sequenced from a random clone library (Burger et al. 2007). For mitochondrial genome assembly we used Phred, Phrap and Consed (Gordon 2003; de la Bastide and McCombie 2007); (http://www.phrap.org/). Mitochondrial genes and introns were identified using automated procedures (MFannot, N. Beck and BFL unpublished; RNAweasel, (Lang, Laforest, and Burger 2007)), followed by manual curation of the results.

Dataset construction

A previously published alignment of nuclear-encoded proteins (Rodriguez-Ezpeleta et al. 2007a) was used for adding the new *Nuclearia* cDNA sequences generated in our lab, plus extra sequences available from GenBank (a taxonomic broad dataset containing 50 eukaryotes will be referred to as the ‘eukaryotic dataset’; another one containing 26 fungal species plus the two *Nuclearia* species as ‘fungal dataset’) using MUST (Philippe 1993) and FORTY (Denis Baurain and HP, unpublished). The number of species has been limited (to allow phylogenomic analyses within reasonable time frames), but only in well-sampled phylogenetic groups of undisputed phylogenetic affinity. Species that were not included are either fast-evolving and/or are incompletely sequenced. Other procedures for dataset construction, in particular the elimination of paralogous proteins, have been described previously (Roure, Rodriguez-Ezpeleta, and Philippe 2007). Within opisthokonts, major lineages had to be represented by at least two distant species, and the extremely fast-evolving Microsporidia were excluded, as these are known to introduce phylogenetic artifacts and an overall reduction of phylogenetic resolution (at an extreme leading to misplacement of species; e.g., (Hirt et al. 1999; Brinkmann et al. 2005)). Sampling within the protist outgroup of the eukaryotic dataset is also not comprehensive (Stramenopila, Alveolata, and Euglenozoa) and limited to slow-evolving representatives of major eukaryotic lineages. The final eukaryotic dataset contains 118 proteins (24 439 amino acid positions) and the fungal dataset 150 proteins (40 925 amino acid positions). Proteins included in the nuclear datasets are listed in supplemental Tables S1 and S2.

For a dataset of mitochondrial proteins, 13 ubiquitous genes (*cox1, 2, 3, cob, atp6, 9,* and *nad1, 2, 3, 4, 4L, 5, 6*) were selected. Muscle ((Edgar 2004)), Gblocks ((Talavera and Castresana 2007)) and an application developed in-house (mams) were used for automatic protein alignment, removal of ambiguous regions and concatenation. The final dataset contains 56 taxa and 2 655 amino acid positions.

Phylogenetic analysis

Phylogenetic analyses were performed at the amino acid (aa) level using methods that are known to be least sensitive to LBA artifacts ((Lartillot and Philippe 2004; Rodriguez-Ezpeleta et al. 2007b; Lartillot and Philippe 2008), and references therein). The concatenated protein datasets were analyzed either by Bayesian inference (BI, PhyloBayes (Lartillot and Philippe 2004)) with the CAT+ Γ model and four discrete gamma categories, or by maximum likelihood (ML, RAxML (Stamatakis 2006) with the WAG+ Γ model and four discrete categories. BI analyses using the CAT model have been shown to be more reliable than ML, due to the application of a more realistic evolutionary model. ML analyses were essentially performed to identify differences in topology, pinpointing problematic parts of the tree for which addition of new data would be in order (i.e., preferentially genome sequences from slowly-evolving species, and those that are expected to break long internal branches at questionable tree topologies).

In case of BI and the eukaryotic dataset (values for the fungal dataset in brackets), chains were run for 3000 (1000) cycles, and the first 1500 (500) cycles were removed as burn-in corresponding to approximately 1,200,000 (400,000) generations. Convergence was controlled by running three independent chains, resulting in identical topologies. The reliability of internal branches for both, ML and BI analyses was evaluated based on 100 bootstrap replicates. For BI, we inferred a consensus tree from the posterior tree topologies of replicates.

Likelihood tests of competing tree topologies were also performed. The site-wise likelihood values were estimated using Tree-Puzzle (Schmidt et al. 2002) with the WAG+ Γ model, and p-values for each topology were calculated with CONSEL

(Shimodaira and Hasegawa 2001).

Variable Length Bootstrap analysis

We compared the performance of nuclear and mitochondrial datasets in phylogenetic inference by Variable Length Bootstrap (VLB) analysis (Springer et al. 2001). Sequences of 29 common species were taken from the eukaryotic (24,439 aa positions) and mitochondrial (2,710 aa positions) datasets. From these, two respective series of sub-datasets were constructed by randomly choosing 400, 600, 800, 1 000 … sequence positions. Phylogenetic inferences were then performed using RAxML with the WAG+Γ model and four discrete categories, after which the BS values for the grouping of nucleariids and Fungi were recorded.

Authors' contributions

ETS and LF constructed and sequenced the two Nuclearia EST libraries; YL, HB, HP and BFL conducted phylogenetic analyses. All authors participated in writing, read and approved the manuscript.

Acknowledgments

We thank Mary L. Berbee (University of British Columbia, Canada) and Rytas Vilgalys (Duke University) for providing *Mortierella* and *Batrachochytrium* cDNA libraries for sequencing, the NHGRI/Broad Institute for access to several new genome sequences (*Allomyces*, *Mortierella*, *Spizellomyces* and *Capsaspora*) and the Canadian Research Chair

Program (BFL, HP), the Canadian Institute of Health Research (BFL) and the ‘Bourses d’Excellence biT’ (CIHR; YL) for salary and interaction support.

Table 1. Comparison of alternative tree topologies with AU and wSH tests.

Log likelihood differences and AU and wSH p values of top-ranking trees are listed.

Rank	**Tree topology**	**ΔlnL**	**AU**	**wSH**
1	**Best tree** (see Figure 1)	0	**1.000**	**1.000**
2	*Nuclearia* sister of Holozoa	187.9	0	0
3	*Nuclearia* sister of Opisthokonta	237.4	0	0
4	*Nuclearia* sister of Asco- + Basidio- + Zygomycetes	418.2	0	0
5	*Nuclearia* sister of *Capsaspora* + *Amoebidium* + *Sphaeroforma*	478.2	0	0
6	*Nuclearia* sister of Metazoa + *Monosiga*	495.3	0	0
7	*Nuclearia* sister of *Allomyces*	511.1	0	0
8	*Nuclearia* sister of *Spizellomyces*	513.7	0	0
9	*Nuclearia* sister of *Capsaspora*	534.3	0	0
10	*Nuclearia* sister of *Amoebidium* + *Sphaeroforma*	561.2	0	0
11	*Nuclearia* sister of Amoebozoa	621.2	0.002	0
12	*Nuclearia* sister of Opisthokonta + Amoebozoa	626.8	0.002	0
13	*Nuclearia* sister of Asco- + Basidiomycetes	704.5	0	0
14	*Nuclearia* sister of *Monosiga*	727.5	0	0
15	*Nuclearia* sister of Metazoa	738.9	0	0

Figure 1: Tree of eukaryotes based on eukaryotic dataset. Trees were inferred with PhyloBayes and rooted following a previous suggestion (Philippe et al. 2000a; Stechmann and Cavalier-Smith 2002). The values at branches indicate bootstrap support (BS) values (upper value, BI/CAT model; lower value ML/WAG model). Values below 60% are indicated by a hyphen; when BS values are equal only one is indicated. The posterior probability values using PhyloBayes are 1.0 for all except two branches (0.98 for the branch uniting Viridiplantae and Haptophyceae; 0.90 for the clade indicted by *). The analyses using ML (RAxML, WAG+Gamma; four categories, Supplementary Material, Figure S1) support the alternative grouping of Malawimonadozoa and JEH with a BS of 77%. Other minor differences include Plantae relationships and the placement of Haptophyceae, which receive no solid support in both BI and ML analyses.

Figure 2: Phylogeny inferred from the mitochondrial dataset. For details on figure description, evolutionary models and phylogenetic methods, see legend of Fig. 1. Note that as already noted in a previous publication (Ruiz-Trillo et al. 2008), the phylogenetic position of *Capsaspora* with mitochondrial data differs from that with nuclear data (Fig. 1). We attribute this inconsistency to the limited availability of mtDNA sequences from *Capsaspora* relatives, and a strong LBA artifact introduced by the fast-evolving Bilateria in concert with *Trichoplax*. Further, the placements of *Cryptococcus* and *Ustilago* differ (although without significant support) from those with nuclear data (see Fig. 4), although results with the much larger nuclear dataset are more likely to be correct.

Figure 3: VLB analysis. Relationship between the number of sequence positions and bootstrap support for Fungi+Nucleariida, with nuclear and mitochondrial datasets.

Figure 4: Fungal phylogeny with nuclear data, using Nucleariida as the outgroup. For details on figure description, evolutionary models and phylogenetic methods, see legend of Figure 1. Note that the phylogenetic position of Blastocladiomycota is unstable, differing between ML *versus* BI analyses (we consider the latter to be more reliable).

Figure 1

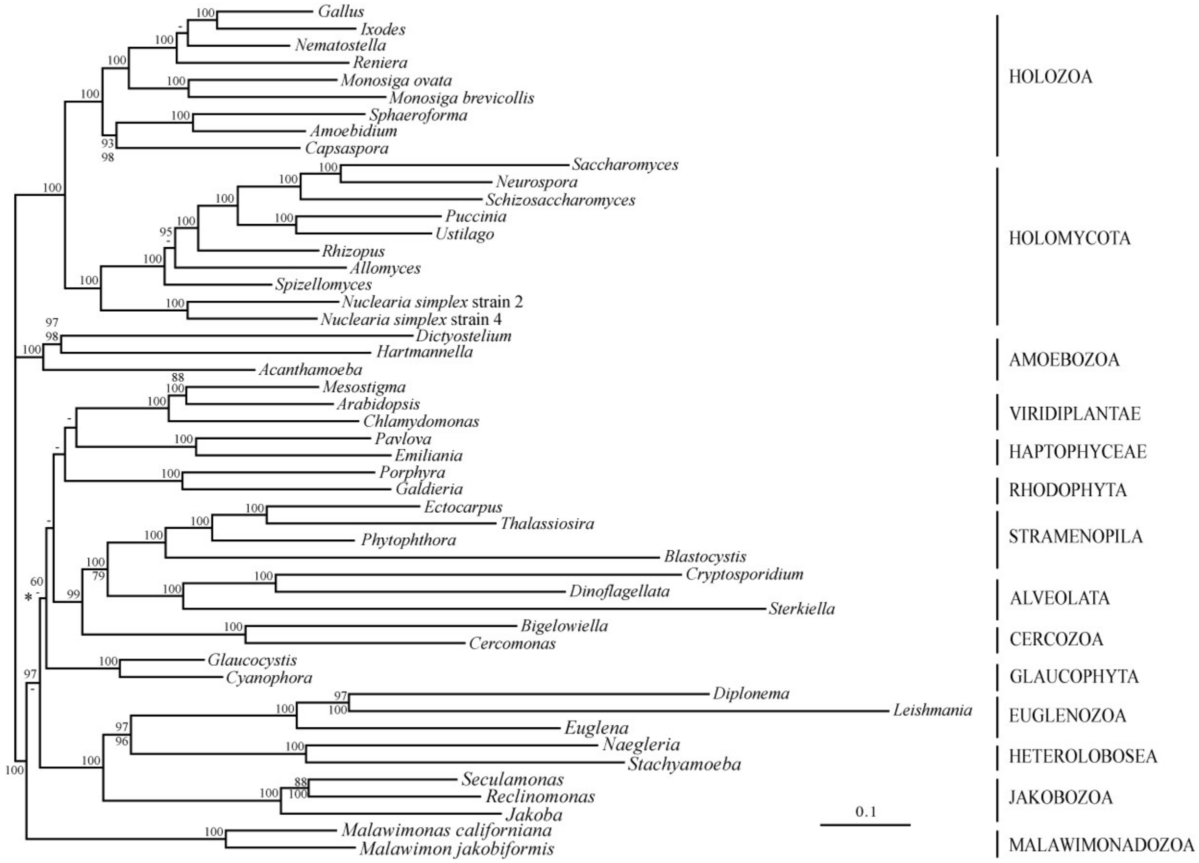

100
Gallus
-
Ixodes
100
Nematostella
100
Reniera
100
Monosiga ovata
100
Monosiga brevicollis
100
Sphaeroforma
Amoebidium
93
98
Capsaspora
100
100
Saccharomyces
100
Neurospora
Schizosaccharomyces
100
100
Puccinia
95
-
Ustilago
Rhizopus
100
Allomyces
100
Spizellomyces
100
Nuclearia simplex strain 2
Nuclearia simplex strain 4
97
98
Dictyostelium
100
Hartmannella
Acanthamoeba
88
100
Mesostigma
100
Arabidopsis
-
Chlamydomonas
-
100
Pavlova
Emiliania
100
Porphyra
Galdieria
100
Ectocarpus
100
Thalassiosira
-
100
Phytophthora
Blastocystis
100
79
100
Cryptosporidium
60
*
-
99
100
Dinoflagellata
Sterkiella
100
Bigelowiella
Cercomonas
100
Glaucocystis
97
-
Cyanophora
97
Diplonema
100
100
Leishmania
97
96
Euglena
100
Naegleria
100
Stachyamoeba
100
88
Seculamonas
100
100
Reclinomonas
Jakoba
0.1
100
Malawimonas californiana
Malawimon jakobiformis
HOLOZOA
HOLOMYCOTA
AMOEBOZOA
VIRIDIPLANTAE
HAPTOPHYCEAE
RHODOPHYTA
STRAMENOPILA
ALVEOLATA
CERCOZOA
GLAUCOPHYTA
EUGLENOZOA
HETEROLOBOSEA
JAKOBOZOA
MALAWIMONADOZOA

Figure 2

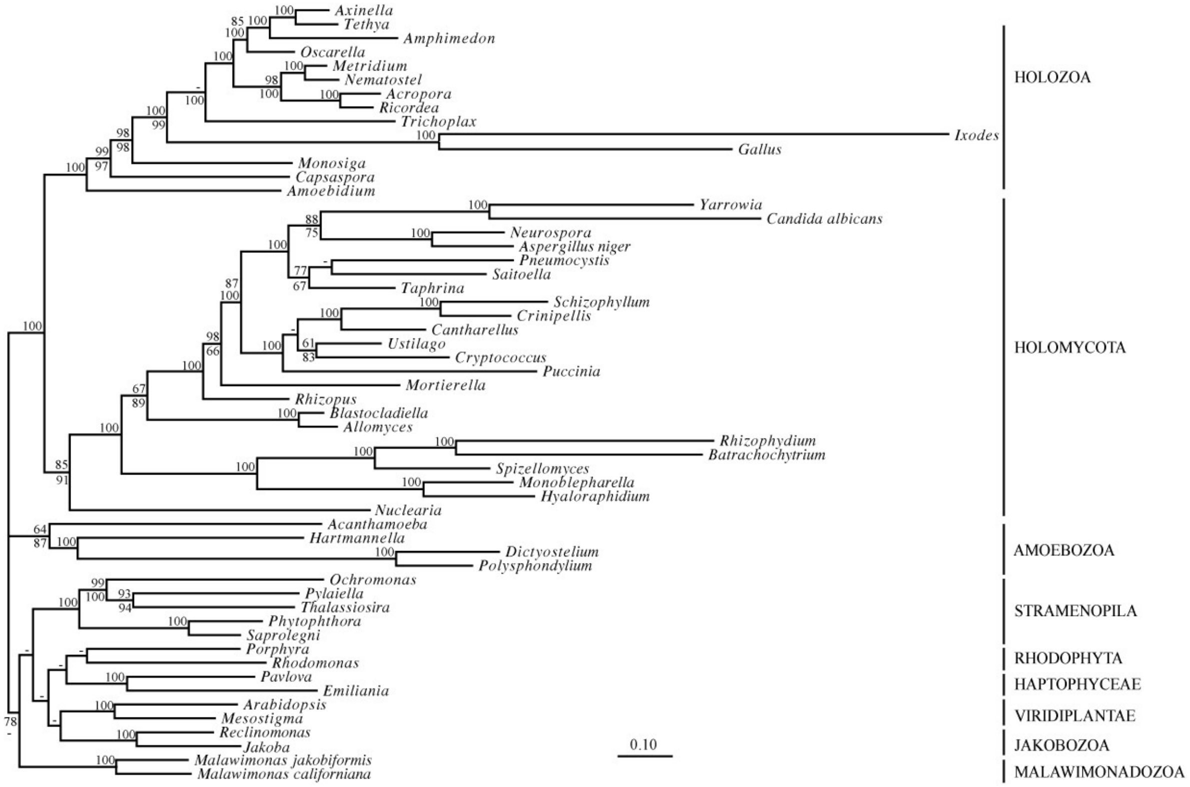

100
Axinella
85 100
Tethya
100
Amphimedon
Oscarella
100
100
Metridium
98
Nematostel
-
100
100
Acropora
100
Ricordea
100
Trichoplax
99
100
Ixodes
98
Gallus
99
98
100
97
Monosiga
Capsaspora
Amoebidium
100
Yarrowia
88
Candida albicans
75
100
Neurospora
100
Aspergillus niger
-
Pneumocystis
77
Saitoella
67
Taphrina
87
100
100
Schizophyllum
100
Crinipellis
Cantharellus
98
-
66
61
Ustilago
100
100
83
Cryptococcus
Puccinia
Mortierella
67
Rhizopus
89
100
Blastocladiella
Allomyces
100
100
Rhizophydium
100
Batrachochytrium
100
Spizellomyces
85
100
Monoblepharella
91
Hyaloraphidium
Nuclearia
64
Acanthamoeba
87
100
Hartmannella
100
Dictyostelium
Polysphondylium
99
Ochromonas
100
93
Pylaiella
100
94
Thalassiosira
100
Phytophthora
Saprolegni
-
-
Porphyra
Rhodomonas
-
100
Pavlova
Emiliania
-
100
Arabidopsis
-
Mesostigma
78
100
Reclinomonas
-
Jakoba
100
Malawimonas jakobiformis
Malawimonas californiana
0.10
HOLOZOA
HOLOMYCOTA
AMOEBOZOA
STRAMENOPILA
RHODOPHYTA
HAPTOPHYCEAE
VIRIDIPLANTAE
JAKOBOZOA
MALAWIMONADOZOA

Figure 3

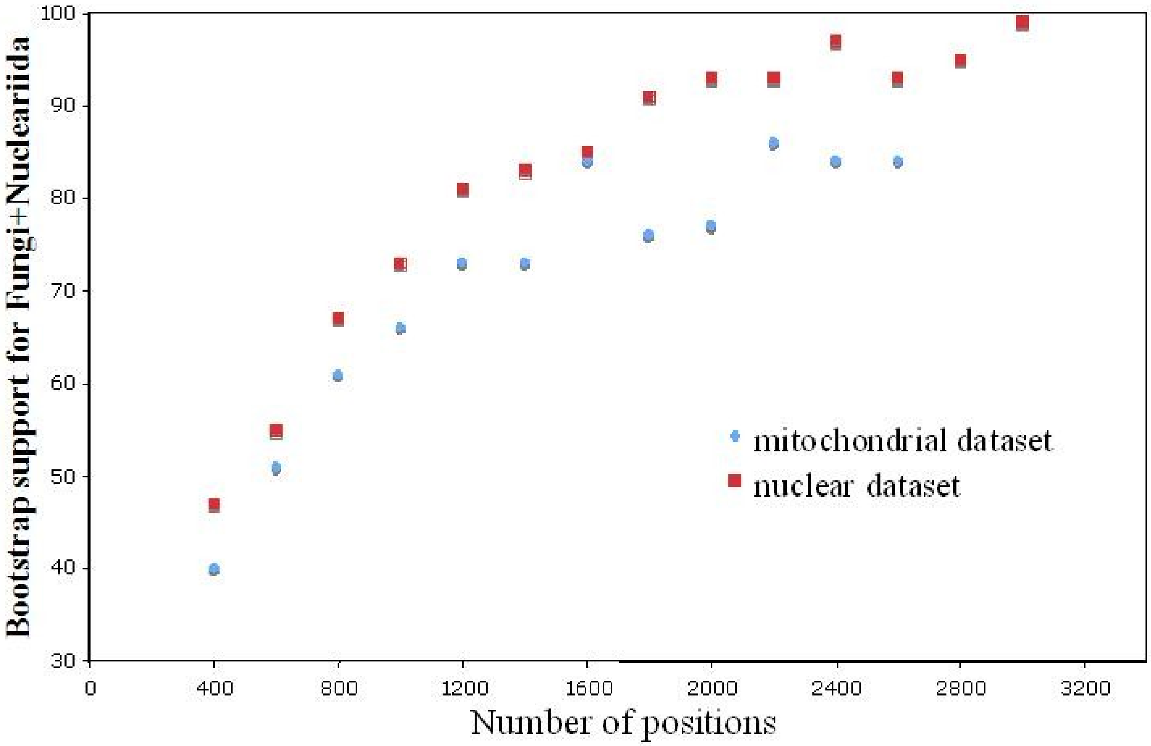
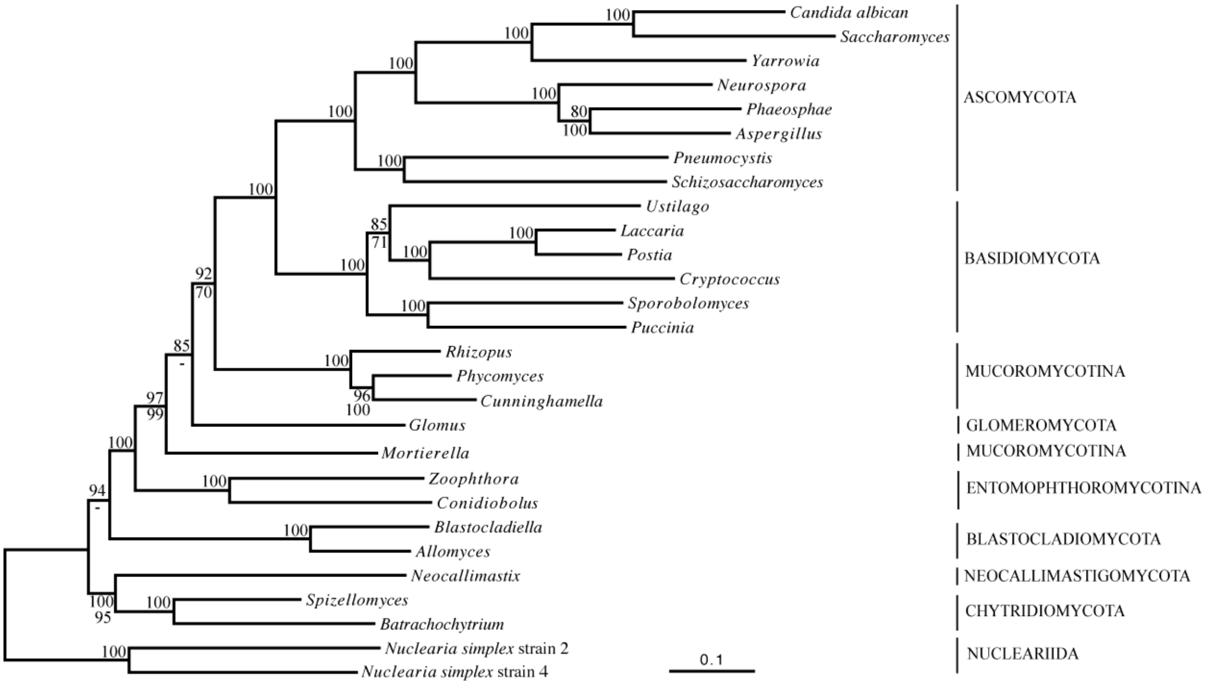

Bootstrap support for Fungi+Nucleariida
100
90
80
70
60
50
40
30
0
400
800
1200
1600
2000
2400
2800
3200
Number of positions
● mitochondrial dataset
■ nuclear dataset

Figure 4


100
Candida albican
Saccharomyces
100
Yarrowia
100
Neurospora
100
80
Phaeosphae
100
Aspergillus
100
ASCOMYCOTA
100
Pneumocystis
Schizosaccharomyces
100
Ustilago
85
71
100
Laccaria
Postia
100
Cryptococcus
BASIDIOMYCOTA
92
70
100
Sporobolomyces
Puccinia
85
-
100
Rhizopus
Phycomyces
96
100
Cunninghamella
MUCOROMYCOTINA
97
99
Glomus
GLOMEROMYCOTA
100
Mortierella
MUCOROMYCOTINA
100
Zoophthora
Conidiobolus
ENTOMOPHTHOROMYCOTINA
94
-
100
Blastocladiella
Allomyces
BLASTOCLADIOMYCOTA
Neocallimastix
NEOCALLIMASTIGOMYCOTA
100
95
100
Spizellomyces
Batrachochytrium
CHYTRIDIOMYCOTA
100
Nuclearia simplex strain 2
Nuclearia simplex strain 4
NUCLEARIIDA
0.1

References

Adl, S. M., A. G. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, T. Y. James, S. Karpov, P. Kugrens, J. Krug, C. E. Lane, L. A. Lewis, J. Lodge, D. H. Lynn, D. G. Mann, R. M. McCourt, L. Mendoza, O. Moestrup, S. E. Mozley-Standridge, T. A. Nerad, C. A. Shearer, A. V. Smirnov, F. W. Spiegel, and M. F. Taylor. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399-451.

Amaral-Zettler, L., T. A. Nerad, C. J. O'Kelly, and M. L. Sogin. 2001. The nucleariid amoebae: more protists at the animal-fungal boundary. J Eukaryot Microbiol 48:293-297.

Berbee, M., and J. Taylor. 2000. The Mycota. Pp. 229-246 *in* E. McLaughlin, E. McLaughlin, and P. Lemke, eds. Fungal Molecular Evolution: Gene trees and geological time. Springer Verlag, New York.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-757.

Burger, G., D. V. Lavrov, L. Forget, and B. F. Lang. 2007. Sequencing complete mitochondrial and plastid genomes. Nat Protoc 2:603-614.

Burki, F., K. Shalchian-Tabrizi, and J. Pawlowski. 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol Lett 4:366-369.

Cavalier-Smith, T. 1987. The origin of eukaryotic and archaebacterial cells. Ann N Y Acad Sci 503:17-54.

Cavalier-Smith, T., and E. E. Chao. 2003. Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J Mol Evol 56:540-563.

de la Bastide, M., and W. R. McCombie. 2007. Assembling genomic DNA sequences with PHRAP. Curr Protoc Bioinformatics Chapter 11:Unit11 14.

Dellaporta, S. L., A. Xu, S. Sagasser, W. Jakob, M. A. Moreno, L. W. Buss, and B. Schierwater. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A 103:8751-8756.

Dykova, I., M. V. I. Fiala, and B. M. H. Peckov. 2003. Nuclearia pattersoni sp. n. (Filosea), a New Species of Amphizoic Amoeba Isolated from Gills of Roach (Rutilus rutilus), and its Rickettsial Endosymbiont. Folia Parasitologica 50:161-170.

Dykova, I., and J. Lom. 2004. Advances in the knowledge of amphizoic amoebae infecting fish. Folia Parasitol (Praha) 51:81-97.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool 27:27-33.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst Biol 53:485-495.

Gordon, D. 2003. Viewing and editing assembled sequences using Consed. Curr Protoc Bioinformatics Chapter 11:Unit11 12.

Hackett, J. D., H. S. Yoon, S. Li, A. Reyes-Prieto, S. E. Rummele, and D. Bhattacharya. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol Biol Evol 24:1702-1713.

Hampl, V., L. Hug, J. W. Leigh, J. B. Dacks, B. F. Lang, A. G. Simpson, and A. J. Roger. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A 106:3859-3864.

Hertel, L. A., C. J. Bayne, and E. S. Loker. 2002. The symbiont *Capsaspora owczarzaki*, nov. gen. nov. sp., isolated from three strains of the pulmonate snail *Biomphalaria glabrata* is related to members of the Mesomycetozoea. Int J Parasitol 32:1183-1191.

Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res 111:509-547.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580-585.

Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy, and T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618-622.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K.

Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006a. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443:818-822.

James, T. Y., P. M. Letcher, J. E. Longcore, S. E. Mozley-Standridge, D. Porter, M. J. Powell, G. W. Griffith, and R. Vilgalys. 2006b. A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). Mycologia 98:860-871.

James, T. Y., D. Porter, C. A. Leander, R. Vilgalys, and J. E. Longcore. 2000. Molecular phylogenetics of the Chytridiomycota supports the utility of ultrastructural data in chytrid systematics. Can J Bot 78:226-350.

Jimenez-Guri, E., H. Philippe, B. Okamura, and P. W. Holland. 2007. *Buddenbrockia* is a cnidarian worm. Science 317:116-118.

Kaiser, D. 2001. Building a multicellular organism. Annu Rev Genet 35:103-123.

Keeling, P. J. 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet Biol 38:298-309.

Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray. 2005. The tree of eukaryotes. Trends in Ecology & Evolution 20:670-676.

Koski, L. B., M. W. Gray, B. F. Lang, and G. Burger. 2005. AutoFACT: An Automatic Functional Annotation and Classification Tool. BMC Bioinformatics 6:151.

Lang, B. F., and G. Burger. 2007. Purification of mitochondrial and plastid DNA. Nat Protoc 2:652-660.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr Biol 12:1773-1778.

Lara, E., D. Moreira, and P. Lopez-Garcia. 2009. The Environmental Clade LKM11 and Rozella Form the Deepest Branching Clade of Fungi. Protist:Epub ahead of print

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Lartillot, N., and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci 363:1463-1472.

Lee, J., and J. P. Young. 2009. The mitochondrial genome sequence of the arbuscular mycorrhizal fungus Glomus intraradices isolate 494 and implications for the phylogenetic placement of Glomus. New Phytol, in press.

Lee, S. C., N. Corradi, E. J. Byrnes, 3rd, S. Torres-Martinez, F. S. Dietrich, P. J. Keeling, and J. Heitman. 2008. Microsporidia evolved from ancestral sexual fungi. Curr Biol 18:1675-1679.

Liu, Y., J. W. Leigh, H. Brinkmann, M. T. Cushion, N. Rodriguez-Ezpeleta, H. Philippe, and B. F. Lang. 2009. Phylogenomic analyses support the monophyly of Taphrinomycotina, including *Schizosaccharomyces* fission yeasts. Mol Biol Evol 26:27-34.

Liu, Y. J., M. C. Hodson, and B. D. Hall. 2006. Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. BMC Evol Biol 6:74.

Medina, M. C., Allen G.; Taylor, John W.; Valentine, James W.; Lipps, Jere H.; Amaral-Zettler, Linda; Sogin, Mitchell L. 2003. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. International Journal of Astrobiology 2:203-211.

Nikolaev, S. I., C. Berney, J. F. Fahrni, I. Bolivar, S. Polet, A. P. Mylnikov, V. V. Aleshin, N. B. Petrov, and J. Pawlowski. 2004. The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc Natl Acad Sci U S A 101:8066-8071.

O'Brien, E. A., L. B. Koski, Y. Zhang, L. Yang, E. Wang, M. W. Gray, G. Burger, and B. F. Lang. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res 35:D445-451.

Patterson, D. 1984. The genus Nuclearia (Sarcodina, Filosea): species composition and characteristics of the taxa. Archiv fuer Protistenkunde 128:127-139.

Patterson, D. J. 1999. The diversity of eukaryotes. Am Nat 154:s96-s124.

Philippe, H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res 21:5264-5272.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci 267:1213-1221.

Redecker, D. 2002. New views on fungal evolution based on DNA markers and the fossil record. Res Microbiol 153:125-130.

Rodriguez-Ezpeleta, N., H. Brinkmann, G. Burger, A. J. Roger, M. W. Gray, H. Philippe, and B. F. Lang. 2007a. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol 17:1420-1425.

Rodriguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe. 2007b. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389-399.

Rodriguez-Ezpeleta, N., S. Teijeiro, L. Forget, G. Burger, and B. F. Lang. 2009. 3. Generation of cDNA libraries: Protists and Fungi. *in* J. Parkinson, ed. Methods in Molecular Biology: Expressed Sequence Tags (ESTs). Humana Press, Totowa, NJ.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol 7 Suppl 1:S2.

Ruiz-Trillo, I., G. Burger, P. W. Holland, N. King, B. F. Lang, A. J. Roger, and M. W. Gray. 2007. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet 23:113-118.

Ruiz-Trillo, I., C. E. Lane, J. M. Archibald, and A. J. Roger. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J Eukaryot Microbiol 53:379-384.

Ruiz-Trillo, I., A. J. Roger, G. Burger, M. W. Gray, and B. F. Lang. 2008. A phylogenomic investigation into the origin of metazoa. Mol Biol Evol 25:664-672.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Seif, E., J. Leigh, Y. Liu, I. Roewer, L. Forget, and B. F. Lang. 2005. Comparative mitochondrial genomics in zygomycetes: bacteria-like RNase P RNAs, mobile elements and a close source of the group I intron invasion in angiosperms. Nucleic Acids Res 33:734-744.

Shalchian-Tabrizi, K., M. A. Minge, M. Espelund, R. Orr, T. Ruden, K. S. Jakobsen, and T. Cavalier-Smith. 2008. Multigene phylogeny of choanozoa and the origin of animals. PLoS One 3:e2098.

Shen, Y.-Q., E. A. O'Brien, L. Koski, B. F. Lang, and G. Burger. 2009. 11. EST Databases and Web Tools for EST Projects *in* J. Parkinson, ed. Methods in Molecular Biology: Expressed Sequence Tags (ESTs). Humana Press, Totowa, NJ.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol 51:492-508.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-1247.

Spatafora, J. W., G. H. Sung, D. Johnson, C. Hesse, B. O'Rourke, M. Serdani, R. Spotts, F. Lutzoni, V. Hofstetter, J. Miadlikowska, V. Reeb, C. Gueidan, E. Fraker, T.

Lumbsch, R. Lucking, I. Schmitt, K. Hosaka, A. Aptroot, C. Roux, A. N. Miller, D. M. Geiser, J. Hafellner, G. Hestmark, A. E. Arnold, B. Budel, A. Rauhut, D. Hewitt, W. A. Untereiner, M. S. Cole, C. Scheidegger, M. Schultz, H. Sipman, and C. L. Schoch. 2006. A five-gene phylogeny of Pezizomycotina. Mycologia 98:1018-1028.

Springer, M. S., R. W. DeBry, C. Douady, H. M. Amrine, O. Madsen, W. W. de Jong, and M. J. Stanhope. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol Biol Evol 18:132-143.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stechmann, A., and T. Cavalier-Smith. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89-91.

Steenkamp, E. T., J. Wright, and S. L. Baldauf. 2006. The protistan origins of animals and fungi. Mol Biol Evol 23:93-106.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.

Tanabe, Y., M. M. Watanabe, and J. Sugiyama. 2005. Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): Insights from molecular phylogenetics. J Gen Appl Microbiol 51:267-276.

Taylor, J., J. Spatafora, K. O'Donnell, F. Lutzoni, T. James, D. Hibbett, D. Geiser, T. Bruns, and M. Blackwell. 2004. The Fungi. Pp. 171-194 *in* M. J. D. Joel Cracraft, ed. Assembling the Tree of Life. Oxford University Press, New York.

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

SUPPLEMENTAL MATERIAL

Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support

Yu Liu[1,3]*, Emma T. Steenkamp[2]*, Henner Brinkmann[1], Lise Forget[1], Hervé Philippe[1] and B. Franz Lang[1]

[1] *Robert Cedergren Centre, Département de biochimie, Université de Montréal, Montréal, Québec, Canada.*

[2] *Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa.*

[3] Present address*: Donnelly Centre for Cellular and Bio-molecular Research, Department of Molecular Genetics, University of Toronto, 160 College Street, Toronto, ON, Canada, M5S 3E1*

Table S1: Proteins in the Eukaryotic Dataset

arp23	cct-A	cct-B	cct-D	cct-E	cct-G	cct-N	cct-T
cct-Z	cpn60mt	ef1-EF1	ef1-RF3	ef2-EF2	fibri	grc5	if1a
if2b	if2g	if6	ino1	l12e-A	l12e-C	l12e-D	nsf1-C
nsf1-G	nsf1-I	nsf1-J	nsf1-K	nsf1-L	nsf1-M	nsf2-A	psma-A
psma-B	psma-C	psma-D	psma-E	psma-F	psma-G	psmb-H	psmb-I
psmb-J	psmb-K	psmb-L	psmb-M	psmb-N	rad51-A	rf1	rpl1
rpl11b	rpl12b	rpl13	rpl14a	rpl15a	rpl16b	rpl17	rpl18
rpl19a	rpl2	rpl20	rpl21	rpl22	rpl23a	rpl24-A	rpl25
rpl26	rpl27	rpl3	rpl30	rpl31	rpl32	rpl33a	rpl34
rpl35	rpl37a	rpl38	rpl39	rpl42	rpl43b	rpl4B	rpl5
rpl6	rpl7-A	rpl9	rpp0	rps1	rps10	rps11	rps13a
rps14	rps15	rps16	rps17	rps18	rps19	rps2	rps20
rps22a	rps23	rps25	rps26	rps27	rps28a	rps29	rps3
rps4	rps5	rps6	rps8	sap40	srp54	srs	suca
vata	vatb	vatc	Vate	w09c			

Table S2: Proteins in the Fungal Dataset

arc20	arp23	cct-A	cct-B	cct-D	cct-E	cct-G	cct-N
cct-T	cct-Z	cpn60mt	crfg	ef1-EF1	ef1-RF3	ef2-EF2	ef2-U5
eif5a	er1	fibri	fpps	grc5	hsp70-E	hsp70mt	hsp90-C
if1a	if2b	if2g	if2p	if4a-a	if4a-b	if6	ino1
l12e-A	l12e-B	l12e-C	l12e-D	mcm-B	mcm-C	mcm-E	mcm-F
nsf1-G	nsf1-J	nsf1-K	nsf1-L	nsf1-M	nsf2-A	pace2-A	pace2-C
psma-A	psma-B	psma-C	psma-D	psma-E	psma-F	psma-G	psmb-H
psmb-J	psmb-K	psmb-L	psmbM	psmb-N	rad23	rad51-A	rf1
rpl1	rpl11b	rpl12b	rpl13	rpl14a	rpl15a	rpl16b	rpl17
rpl18	rpl19a	rpl2	rpl20	rpl21	rpl22	rpl23a	rpl24-A
rpl25	rpl26	rpl27	rpl3	rpl30	rpl31	rpl32	rpl33a
rpl34	rpl35	rpl36	rpl37a	rpl38	rpl39	rpl42	rpl43b
rpl4B	rpl5	rpl6	rpl7-A	rpl9	rpo-A	rpo-B	rpo-C
rpp0	rps1	rps10	rps11	rps13a	rps14	rps15	rps16
rps17	rps18	rps19	rps2	rps20	rps22a	rps23	rps24
rps25	rps26	rps27	rps27a	rps28a	rps29	rps3	rps4
rps5	rps6	rps7	rps8	rps9	sap40	srp54	srs
suca	tfiid	tif2a	vata	vatb	vate	xpb	vatpased
ATP synthase-mt		dihydrolatransacylase-b			ornamtrans-a		
pyrdehydroe1b-mt		sadhchydrolase-E1			vacaatpasepl21-a		

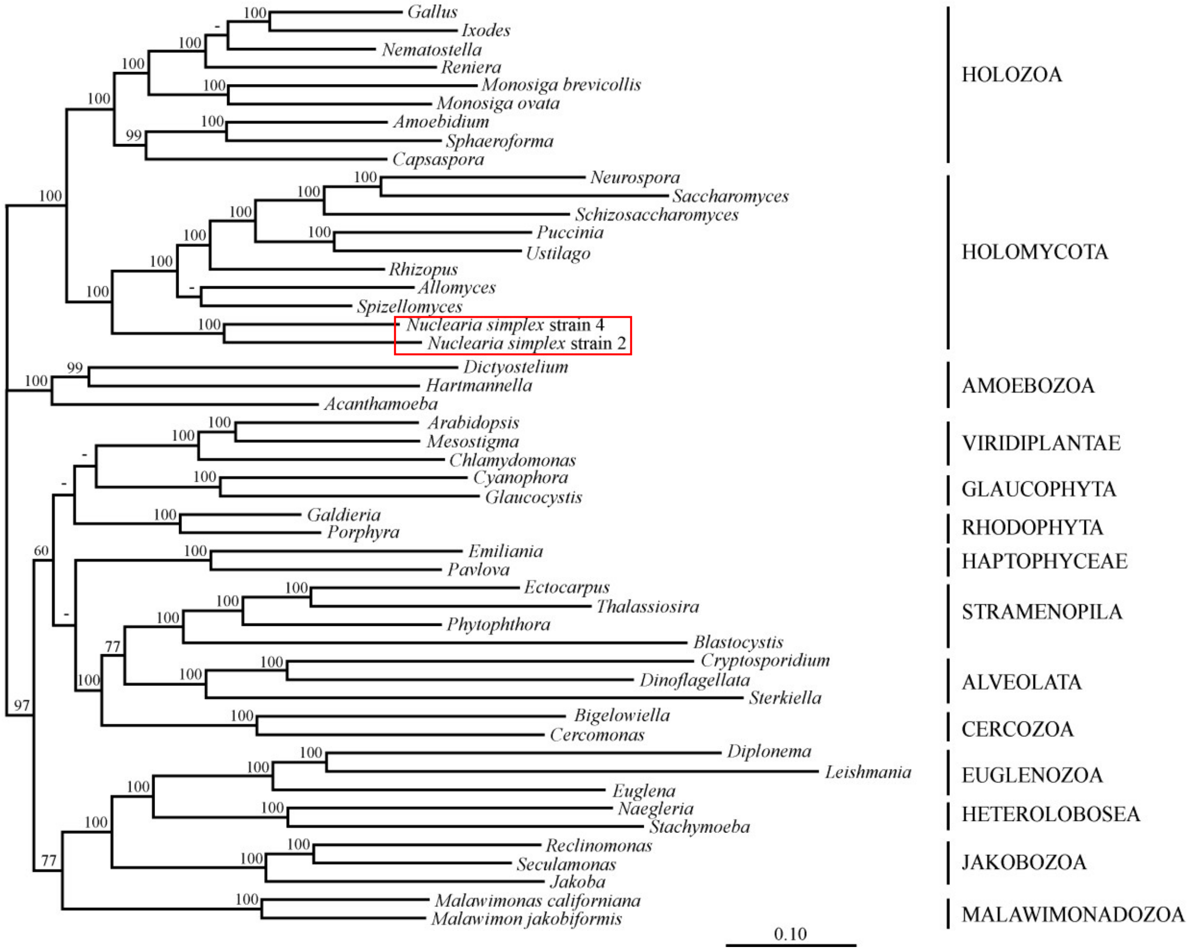


Figure S1: Tree of eukaryotes with the eukaryotic dataset and ML inference. The analyses using RAxML (WAG+Gamma; four categories) (Stamatakis 2006) support the grouping of Malawimonadozoa and JEH with a BS of 77%. Other differences with the tree using BI method (Figure 1) include Plantae relationships, and the placement of Haptophyceae, which receive no support for both BI and ML analyses with this dataset. For more details see legend of Fig. 1.

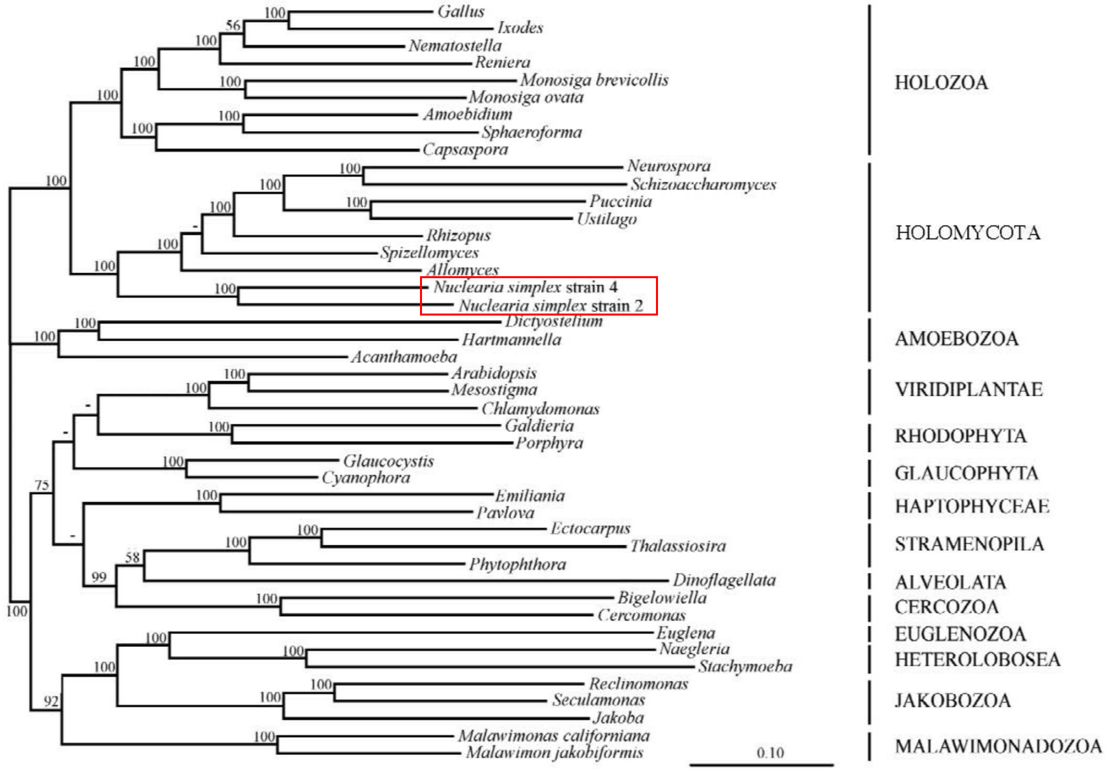


Figure S2: Phylogeny with Eukaryotic Dataset after removing fast-evolving species.

The tree was inferred with ML (RAxML) using the WAG+Gamma model with four categories. Numbers at branches represent support values obtained with 100 bootstrap replicates.

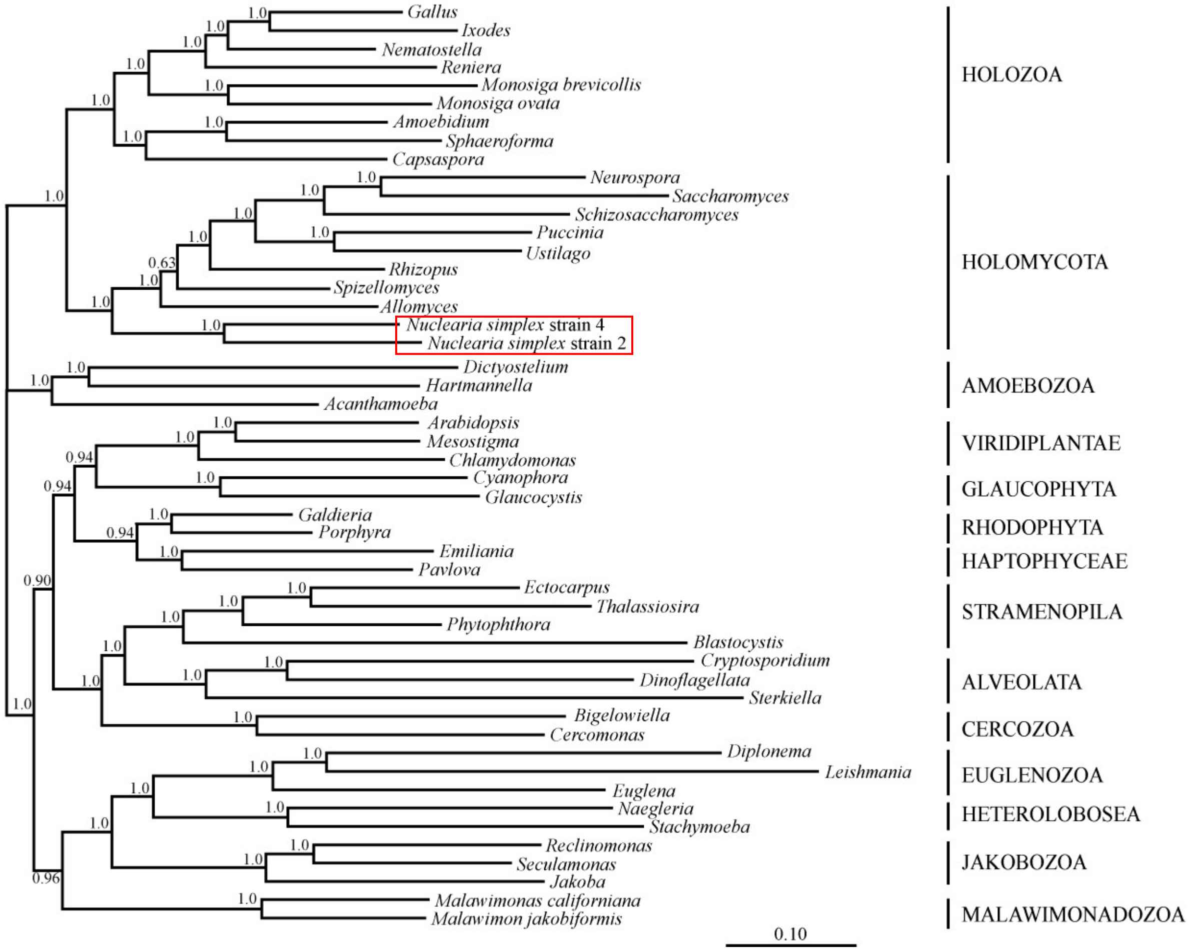


Figure S3: Phylogeny with Eukaryotic Dataset after recoding amino acids into six groups. The amino acids of the Eukaryotic Dataset were recoded into six groups as follows: (1) ASTGP, (2) DNEQ, (3) RKH, (4) MVIL, (5) FYW and, (6) C. This allowed the use of a 6 × 6 general time-reversible rate matrix with free parameters rather than a fixed empirical matrix. Sequence composition and among-site rate variation parameters were also free in the BI analysis, as implemented in P4 (Foster 2004); 50 000 generations, first 10 000 removed as burn-in). Numbers at branches represent PP values.

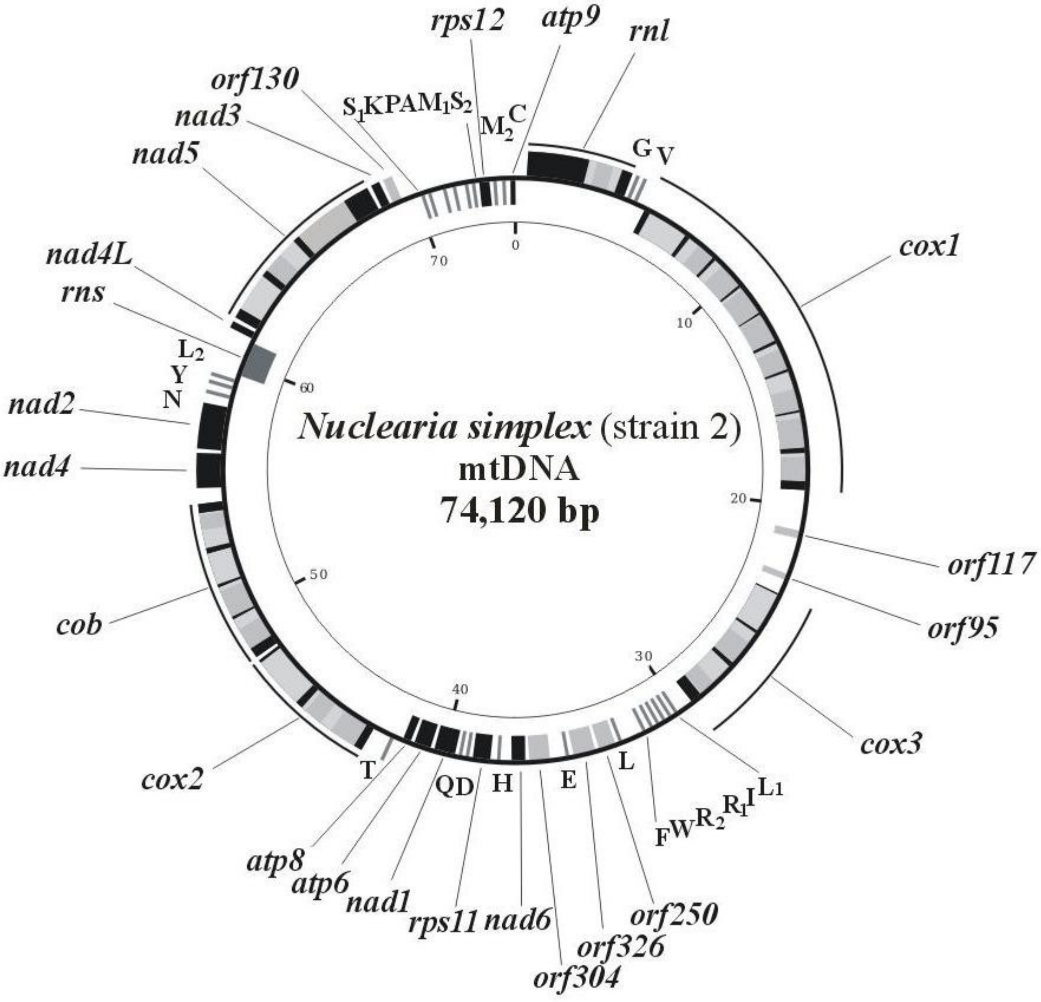


Figure S4: Complete mitochondrial DNA (mtDNA) of *Nuclearia simplex* strain 1552/2. The circular-mapping mtDNA (74,120 bp) is displayed starting with the *rnl* gene (coding for the large subunit rRNA), clockwise in direction of transcription. Black bars, genes or exons; grey bars, introns and intronic ORFs; tRNA genes are named by the one-letter amino acid code.

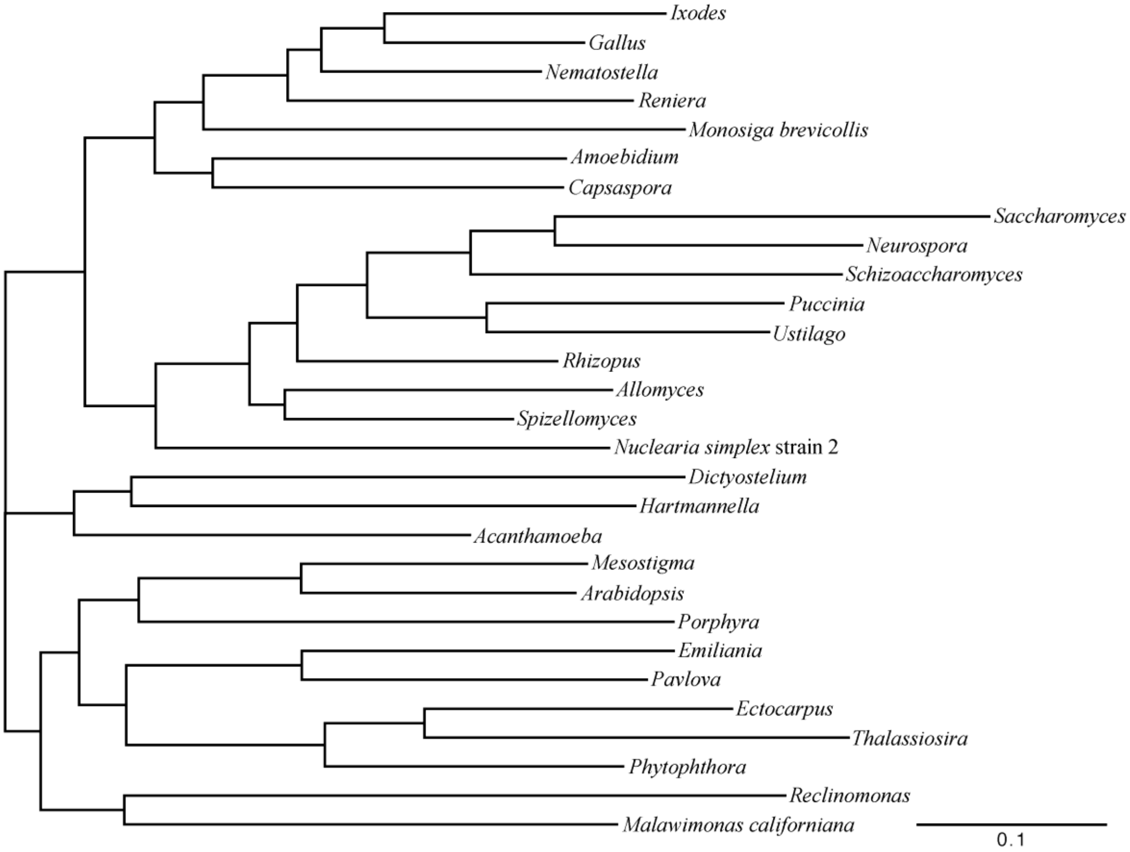


Figure S5: Phylogenetic relationship of the 29 species used for VLB analyses.

The phylogeny was inferred using nuclear data and RAxML with the WAG + Gamma model. In order to minimize missing data in the mitochondrial dataset, we exchanged *Saccharomyces* and *Schizosaccharomyces* that both lost the six mitochondrion-encoded *nad* genes with *Candida albicans* and *Taphrina.*

References

Foster PG: **Modeling compositional heterogeneity**. *Syst Biol* 2004, **53**(3):485-495.

Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics (Oxford, England)* 2006, **22**(21):2688-2690.

Chapter 5 A likelihood ratio test to identify highly heterotachous sites in protein sequences - application to phylogenomic analysis

Yu Liu[1,2]

[1]*Robert Cedergren Centre, Département de biochimie, Université de Montréal, Montréal, Québec, Canada.*

[2]Present address: *Case Center for Proteomics and Bioinformatics, School Of Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, U.S.A.*

Abstract

Long Branch Attraction (LBA) is a common phylogenetic artifact that leads to the grouping of species with elevated evolutionary rates - irrespective of their true phylogenetic position. In these species, a large fraction of fast evolving sequence sites is virtually randomized (over-saturated) without phylogenetically tractable information. Recent studies demonstrate that heterotachy (within-site rate variation) is prevalent in phylogenetic datasets, predominating in fast evolving species. If not accounted for by the evolutionary model, heterotachy may escalate the effect of LBA. Yet, current implementations of heterotachous models are computationally too demanding to be used with large datasets. Until these methods are more developed, removal of highly heterotachous (HH, identified as significant by statistical method) sites in groups of fast-evolving species may provide the most effective alternative to counteract LBA. Here I present a method based on a Likelihood Ratio Test (LRT) that permits progressive elimination of HH sites. In contrast to other methods, I identify and eliminate sequence positions only in the fast-evolving taxa, i.e., data loss due to sequence site removal is limited. Two previously published datasets with known, strong LBA effects are used to demonstrate the potential of our method. When using maximum likelihood (ML) as inference method, removal of HH sites in fast evolving taxa overcomes LBA in both instances, with bootstrap support values for the expected (true) topologies at 97% in one instance, and 90% in the other.

Introduction

LBA, the grouping of species with elevated evolutionary rates irrespective of their true phylogenetic position, is the most-cited and probably most prevalent artifact in phylogenetic analysis. It affects available inference methods without exception, although at varying degrees (Felsenstein 1978a; Hartmann and Golding 1998; Dean and Golding 2000; Lopez, Casane, and Philippe 2002; Susko et al. 2002; Gribaldo et al. 2003; Philippe et al. 2003; Bevan, Lang, and Bryant 2005; Phillips et al. 2006; Bevan, Bryant, and Lang 2007). Understanding causes and consequences of LBA is crucial for overcoming phylogenetic artifacts in the analysis of real sequence datasets.

In his seminal paper, Felsenstein illustrates that phylogenetic methods will be misled by LBA, when evolutionary rate variation across lineages is high enough to effectively randomize sequence character states with respect to their phylogenetic history (Felsenstein 1978a). More recently it was suggested that another rate heterogeneity, heterotachy (Lopez, Casane, and Philippe 2002), significantly strengthens the effect of LBA (Kolaczkowski and Thornton 2004; Philippe et al. 2005b). According to the authors' interpretation, maximum parsimony (MP) would perform better than maximum likelihood (ML) in recovering the correct topology for datasets with high degrees of heterotachous sites. This study prompted several other studies (Gadagkar and Kumar 2005; Gaucher and Miyamoto 2005; Philippe et al. 2005b; Spencer, Susko, and Roger 2005), confirming that heterotachy is indeed much more prevalent than previously thought. Yet, they all disagree that MP outperforms ML in resolving the correct tree topology. More recently, models and methods have been developed to overcome phylogenetic artifacts caused by heterotachy: mixture models have been developed to improve phylogenetic accuracy, where traditional homotachous models fail (Kolaczkowski and Thornton 2008). Yet a major drawback of these methods is computational load, limiting its application to small datasets. In one of the studies, the proportion of heterotachous sites in real datasets varies between 5% and 60% (< 25% in four datasets, 60% in an extreme case (Pagel and Meade 2008)). This implies that homotachous models typically describe a large fraction of sites correctly, i.e. that removal of heterotachous sites would usually leave sufficient information to resolve phylogenetic questions with homotachous models.

Particularly in case of large datasets, data filtering has long been used to avoid artifacts caused by rate heterogeneity (e.g., (Brinkmann and Philippe 1999; Lopez, Forterre, and Philippe 1999; Susko et al. 2002; Pisani 2004; Brinkmann et al. 2005; Roger and Hug 2006; Rodríguez-Ezpeleta N et al. 2007)). These methods eliminate data columns that are not correctly handled by current evolutionary models, indiscriminately across all species. Therefore, a gradual increase of phylogenetic *versus* non-phylogenetic signal comes at a steep cost, ultimately reducing the dataset to a point where insufficient data is left to resolve the tree. For example, the S-F method (Brinkmann and Philippe 1999) removes fast-evolving positions with least reliable evolutionary information, and the H-P method (Lopez, Forterre, and Philippe 1999) eliminates heterotachous sites from alignments. Although both methods work well for given examples, in order to calculate site-wise evolutionary rates, they require *a priori* grouping of monophyletic taxa (at least three species per monophyletic group) for all species presented in the dataset. When less than three species are available for a given group, these methods are not applicable. An alternative method to detecting heterotachous sites is modeling among-site rate variation through a bivariate discrete rate distribution, with a matrix of 25 by 25 categories (Susko et al. 2002). The major limitation of this method is computationally impracticability for large datasets. In this study, I present a method that is suitable for larger datasets and effectively detects heterotachous sites in a well-defined lineage without a requirement for excluding lineages from the dataset that are represented by less than three taxa.

In addition, instead of indiscriminately removing sites column-wise, I do so only for the group of taxa that is likely affected by LBA. Data of other species are untouched so that the overall phylogenetic resolution remains strong. Our method is based on a Likelihood Ratio Test (LRT) of unrooted tree topologies. LRT has been previously applied to detecting heterotachous sites in paralogous proteins (Knudsen and Miyamoto 2001), but this implementation (without evident motive) requires rooted trees to calculate likelihood values. In many cases, a reliable, close outgroup is unavailable for rooting, and a forced introduction of a (distant) outgroup will predictably increase the level of phylogenetic uncertainty and strengthen LBA artifacts rather than reducing them.

Although H_0 and H_1 is not adequately defined in the original paper (Knudsen and Miyamoto 2001), the procedure for simulating LRT statistics is appropriate, closely matching chi-square distribution (see Fig. 2 in (Knudsen and Miyamoto 2001)). Therefore, this LRT approach may be used for identification of evolutionary rate shift among proteins.

Another concern is that the application of LRT site by site will cause a multiple comparison problem (the more hypotheses are tested simultaneously, the higher the probability of obtaining false positive), increasing the probability of a type 1 error (or false positive). The order of LRT statistics is less sensitive to the multiple comparison problems than the P-value. There are procedures that use it to correct the multiple comparisons (Simes 1986). Although the LRT is used site-by site in our method, the relative order of LRT statistics is used to decide the rank of sites that largely contribute to LBA (i.e., the HH sites). This will cause that our procedure is relative robust to the multiple comparison problem.

Two previously published datasets (Brinkmann et al. 2005; Philippe et al. 2005b) are used to demonstrate the potential of our procedure in phylogenetic analysis, using ML as the basic inference method. The analyses of the original datasets are strongly affected by LBA. Our results demonstrate that the LRT method efficiently detects heterotachous sites, and removing them in specified groups effectively overcomes LBA.

Materials and Methods

Datasets

Two previously published datasets are analyzed in this study. The first one (Philippe, Lartillot, and Brinkmann 2005) includes 146 genes from 49 species. Five of these genes were removed, as these are missing in the fast-evolving platyhelminths. This modified dataset has 33,452 amino acid positions (referred to hereafter as the Animal Dataset). The second dataset with 133 genes ((Brinkmann et al. 2005); referred to hereafter as the

Eukaryotic Dataset) comprises 44 species (six Archaea, 33 slowly evolving eukaryotes, and *Encephalitozoon cuniculi*, a member of the fast evolving Microsporidia).

Phylogenetic analysis of Animal Dataset

Because heuristic approaches might miss the correct solution, two different models (methods) for the phylogenetic inference were compared: (i) a separate model (Bapteste et al. 2002); for details see (Philippe, Lartillot, and Brinkmann 2005); and (ii) a conventional model with concatenated data using RAxML and the WAG amino acid replacement matrix and gamma-distributed rates across sites. Bootstrapping (100 replicates) is used to evaluate the support for internal branches.

Phylogenetic analysis of Eukaryotic Dataset

In the original analysis of the Eukaryotic Dataset (Brinkmann et al. 2005), both separate and concatenated models were applied. The ML tree is presented in Figure 4A. In this study, I exclude the concatenated model due to the concern of trapping in a local maximum and apply the separate analysis with a different approach to constrain the tree space (Brinkmann et al. 2005). I define two different sets of constraints: in one set the monophyly of the five main eukaryotic lineages (animal, plant, stramenopiles, alveolates and Fungi; in the case of Fungi, the four main fungal lineages are left unconstrained) and Archaea are constrained respectively, Microsporidia is free to group with any of them. In the second set, Microsporidia is constrained within the fungal lineage, leaving unconstrained its position in

Fungi; other groups are as in the first set of constraints. In total, these two sets of constraints define 76,545 topologies. Then, the procedure of separate model is applied as in (Bapteste et al. 2002).

Procedure for HH site filtering

The goal of this filtering procedure is the identification and elimination of sequence positions that cannot be properly handled by current inference methods, i.e., HH sites. Data elimination is restricted to a predefined subgroup that is known or suspected to be prone to phylogenetic artifacts (e.g., fast evolving groups). The procedure may also be applied to more than one subgroup, but then the identification of HH sites has to be done separately. The procedure contains the following steps:

1) The dataset is divided into two parts: the targeted subgroup (subgroup 1 in Figure 1) and the rest (subgroup 2 in Figure 1);
2) Positions whose evolutionary rates are significantly different between subgroup 1 and 2 are detected by applying LRT on each sequence position (see next section);
3) Positions detected from last step can be divided into two classes: one includes positions whose rate is higher in subgroup1 than subgroup2; other positions. Comparison of the rates with maximum likelihood values for the two subgroups (evolutionary rate can be estimated using ML method, this is done at the same time with the calculation of topology likelihood) identifies the site in the first class, which are HH sites in subgroup 1;

4) HH sites are ranked by their LRT P values, and filtered progressively.

Details for LRT and HH site removal are described in the following.

Likelihood ratio test

LRT is a commonly used statistical test of the goodness-of-fit between two nested models; it provides an objective criterion for selecting among possible models. A complex model is compared to a simpler model, to see if it fits a dataset significantly better. LRT is only valid for comparing hierarchically nested models: the complex model must differ from the simple model only by the addition of one or more parameters.

In this study, it is used to detect sites with significant evolutionary rate differences between a target group (subgroup 1; for details on the procedure see Fig. 1) and all other species. Starting with a given phylogenetic tree, the dataset is split into two subsets, subgroup 1 and 2. Subgroup 1 is the assumed fast-evolving group. The null model (H_0) of LRT states that a given position *i* has the same evolutionary rate in both sub-datasets (subgroup 1 and 2), and the alternative model (H_1) that this position has different rates. For each site *i*, the maximum likelihood values $L_0(i)$ and $L_1(i)$ are calculated under both models, using a pruning algorithm with PAML (Felsenstein 1981; Yang 1997). The significance of differences between the two models is assessed using the following test statistics: $P(i) = -2\log(L_0(i)/L_1(i))$. The test statistics $P(i)$ is asymptotically chi-square-distributed with one degree freedom, and the critical value (c.v.) under this condition is 3.841, at a 5%

significance level. I denote S as the set of sites with significantly different rates of evolution, at the 5% significance level.

Removal of HH sites

The identification and ranking of HH sites is illustrated in Fig. 2; the aim is their specific identification in subgroup 1. Sites with significant rate differences between subgroup 1 and 2 (a set defined as S) are identified by LRT, without determining whether sites of subgroup 1 evolve faster or slower than in subgroup 2. In order to distinguish these two cases, for each site in S, the evolutionary rates are estimated separately for subgroup 1 and 2, by using PAML (Yang 1997): let $\lambda_1(i)$ and $\lambda_2(i)$ be the evolutionary rates that maximize the likelihood of the phylogeny for position *i* of subgroup 1 and subgroup 2 respectively. I then divide the set S into two subsets (S_1 and S_2) based on $\lambda_1(i)$ and $\lambda_2(i)$: S_1 contains sites with $\lambda_1(i) > \lambda_2(i)$, S_2 the remainder. S_1 is the set of HH sites, and the degree of heterotachy can be ranked according to test statistics P(*i*) of LRT. HH sites may then be progressively eliminated. Perl scripts implementing these procedures are available upon request.

Results and Discussion

Animal phylogeny and LBA

Resolving deep animal phylogeny remains challenging due to LBA artifacts, resulting from strong rate heterogeneity among distant animal lineages. Various methods have been used to detect and overcome LBA, and to improve our knowledge of animal evolution. For example, in order to test if with a given dataset nematodes and platyhelminths group due to

LBA (Fig 3A), Philippe et al. (Philippe, Lartillot, and Brinkmann 2005) apply two approaches. They either exclude fast-evolving nematodes and platyhelminths from the analysis, or progressively filter out fast-evolving genes. Under both conditions, nematodes and platyhelminths are no longer attracted; nematodes group with arthropods and platyhelminths with annelids plus mollusks. Clearly, the grouping of nematodes and platyhelminths result from LBA, and removing biased data (faster evolving species or genes) allows recovering the true phylogenetic relationships.

In this study, I address this problem by using an alternative method, removal of HH sites from the fast evolving nematodes and platyhelminths.

Heterotachous sites in nematodes and platyhelminths

Using LRT, I find that 74.7% of sequence sites in nematodes have significant rate differences compared to other species (at a 5% significance level), i.e., they are heterotachous. Further comparisons of evolutionary rates indicate that a large fraction of heterotachous sites evolve significantly faster in nematodes (49.3% of total sequence sites; defined as HH sites). In the case of platyhelminths, 69.2% of sites have significant rate differences with the rest and 51.4% evolve significantly faster. A comparison of HH sites in nematodes and platyhelminths shows that about half (8,231) are common, and half (8,495) are specifically accelerated in either nematodes or platyhelminths. The fraction of common HH sites is expected to contribute most to LBA, in this example.

HH site removal effectively overcomes LBA artifacts

HH sites are progressively removed in steps of 10% of total sites, and resulting datasets are evaluated by phylogenetic analysis (see summaries in **Tables 1** and **2**). Table 1 shows the resulting best ML trees. Table 2 lists bootstrap values for a branch that is most affected by LBA (marked with an asterisk in Tree B of Figure 3). After removal of 20% or more HH sites from nematodes, the ML tree changes from the incorrect Tree A to the expected Tree B (Fig. 3). HH site removal from platyhelminths leads to similar results. The most efficient way is removal from both groups simultaneously: at only 10%, nematodes group correctly with arthropods, and the bootstrap support in separate analysis increases to 97% (85% with RAxML) when all HH sites are removed. In contrast, the fraction of HH sites that are common to species other than nematodes and platyhelminths is only 20%, and their removal does not change the tree topology (result not shown).

The two methods applied in a previous publication (Philippe, Lartillot, and Brinkmann 2005), complete removal of fast-evolving genes or sequence sites, are expected to result in a decrease of overall phylogenetic resolution. In fact, after removing 75 fast-evolving genes (out of 146), the support for deuterostome monophyly drops from 94% to 75% (Figs. 2 and 4 in (Philippe, Lartillot, and Brinkmann 2005)). Because deuterostome sequences remain untouched and the improvement of the ratio of signal and noise by our procedure, the respective support value increases to significant (100%) from 94% (Fig. 3B).

Position of Microsporidia in the eukaryotic tree

Deep level phylogenies are of fundamental importance for understanding the relationship among eukaryotic supergroups and the origin of eukaryotes. Yet, the presence of several fast evolving supergroups (e.g. Fungi, Alveolata) makes the occurrence of LBA very likely, with the consequence that extremely-fast-evolving taxa such as Microsporidia are difficult to place (Philippe et al. 2000a). Although it is widely accepted that Microsporidia are related to Fungi (Baldauf et al. 2000; Keeling, Luker, and Palmer 2000; Philippe et al. 2000a), special measures have to be taken to overcome LBA, and statistical support for this topology remains limited. The precise positioning of Microsporidia within Fungi by phylogenetic analysis remains unresolved.

In the analysis of Eukaryotic groups, a close outgroup is very useful to avoid LBA (Philippe and Laurent 1998). In most cases, it is unavailable, the addition of a distant outgroup (Archaea) is inevitable, which further increases the potential of LBA. In the ML tree from analysis of the Eukaryotic Dataset (Figure 4A), Archaea and Microsporidia form a monophyly that groups with Fungi. This grouping is very likely due to LBA artifact (Cavalier-Smith 1987b; Baldauf and Palmer 1993; Baldauf et al. 2000). In this study, I apply the LRT to identify the HH sites in the distant outgroup, Archaea, and investigate their effects on phylogenetic analyses.

HH sites in Archaea contribute to LBA artifacts

LRT analysis identifies ~ 30% of the archaeal sequence positions as HH sites (at a 5% significant level), and their removal leads to a change of tree topology in which Microsporidia group with Fungi. The bootstrap support for this topology is weak even after removal of 30% of HH sites (45% for the grouping of Microsporidia with Fungi, Figure 4B). The support increases up to 91% when 40% sites are removed, but decreases to 78 % when 50% sites removed, likely because an increasing number of sites with phylogenetic signal are removed from the dataset.

In the original analysis of the Eukaryotic Dataset, fast-evolving Microsporidia proteins are identified and removed progressively from dataset (Brinkmann et al. 2005). Using a properly constraint tree space, I repeat the analysis using their approach, and find that the support for the grouping of Microsporidia with Fungi increases up to only 63% support after the removal of 90% Microsporidia proteins. Due to the difference of two methods, it is impossible to compare the performance of our method with (Brinkmann et al. 2005).

In summary, the application of LRT on Eukaryotic Dataset confirms that Microsporidia has a close relation with Fungi, although with moderate support. More data from other Microsporidia species are needed to resolve its placement with confidence. This example also demonstrates that removal of data which contain no or few phylogenetic signal from distant outgroup can overcome LBA and improve the accuracy of phylogenetic inference.

Acknowledgments

I thank B Franz Lang and Hervé Philippe for providing sequence alignments and for helpful suggestions, and Rachel Bevan, Jessica Leigh, and Nicolas Rodrigue for critical reading of the manuscript. Support from the 'Bourses d'Excellence biT' (CIHR) is acknowledged.

Table 1: Correlation of tree topology and the degree of HH site removal in nematoda and/or platyhelminths (Animal Dataset).

	10%	20%	30%	40%	50%
Nematoda	Tree A*	Tree B*	Tree B	Tree B	Tree B
Platyhelminths	Tree A	Tree B	Tree B	Tree B	Tree B
Both	Tree B	Tree B	Tree B	Tree B	Tree B

* Tree A is the incorrect LBA topology in which nematodas and platyhelminths group together;

* Tree B reflects the expected relationships among animal lineages (for tree topologies see Fig. 3).

Table 2: Variation of bootstrap support for the gorup of Nematoda + Arthropoda with HH site removal in Nematoda and/or Platyhelminths.

	10%		20%		30%		40%		50%	
	Sep.*	Con.*	Sep.	Con.	Sep.	Con.	Sep.	Con.	Sep.	Con.
Nematoda	32.2	14	46.2	25	64.0	42	69.1	59	83.0	66
Platyhelminths	41.4	17	52	31	61.0	26	74.4	46	87.7	67
Both	57.5	21	75.1	37	86.1	44	89.6	55	97.0	85

* Sep, Separate analysis

* Concatenated model using RAxML.

FIGURE LEGENDS

Figure 1: Illustration of LRT to identify heterotachous sites.

For each site, the two models of LRT are: H_0: the two subgroups share the same rate λ; H_1: subgroups have different rates λ_1, λ_2. The likelihood values under the two models are: $L_0 = \max(L_{sub1}(\lambda)L_{sub2}(\lambda)$; $L_1 = \max(L_{sub1}(\lambda_1))\max(L_{sub2}(\lambda_2))$. The test statistics is calculated as: $P(i) = -2(\log L_0 - \log L_1)$. The dataset is divided into two parts: one that fits H_0, and another H_1 that belongs to set S (see Materials and Methods section for further details).

Figure 2: Procedure to identify heterotachous sites of a target group, and to rank them by LRT. For each site of set S from Fig 1, the site-wise rates are estimated and compared between the target group and the rest. The HH sites are those that evolve significantly faster in the target group. They are ranked based on the test statistics of LRT.

Figure 3: Trees of Animal Dataset. The same topology was obtained using either a separate or concatenated WAG+F+G model (RAxML). The values close to internal branches indicate bootstrap support values of the separate (upper) or concatenated (lower) models. When both are 100, only one is indicated; when one is below 75 it is indicated by -. Tree A: ML tree from complete dataset; Tree B: 20% of HH positions are removed from Nematoda. The best tree and bootstrap values for the node with * in cases of other percentages of site removal in Nematoda and/or Platyhelminths are listed in Table 1 and 2.

Figure 4: Analysis of the Eukaryotic Dataset. The tree was inferred with WAG+F+G using separate analysis. Bootstrapping support for internal branches is indicated above branches. The asterisk indicates groups that are constrained in the analysis. Tree A: ML tree from the analysis of original complete dataset; Tree B: ML tree and bootstrap support after removal of HH sites from Archaea. The values at nodes are bootstrapping supports after 50%, 40% and 30% (top to bottom) removal of archaeal sites.

Figure 1

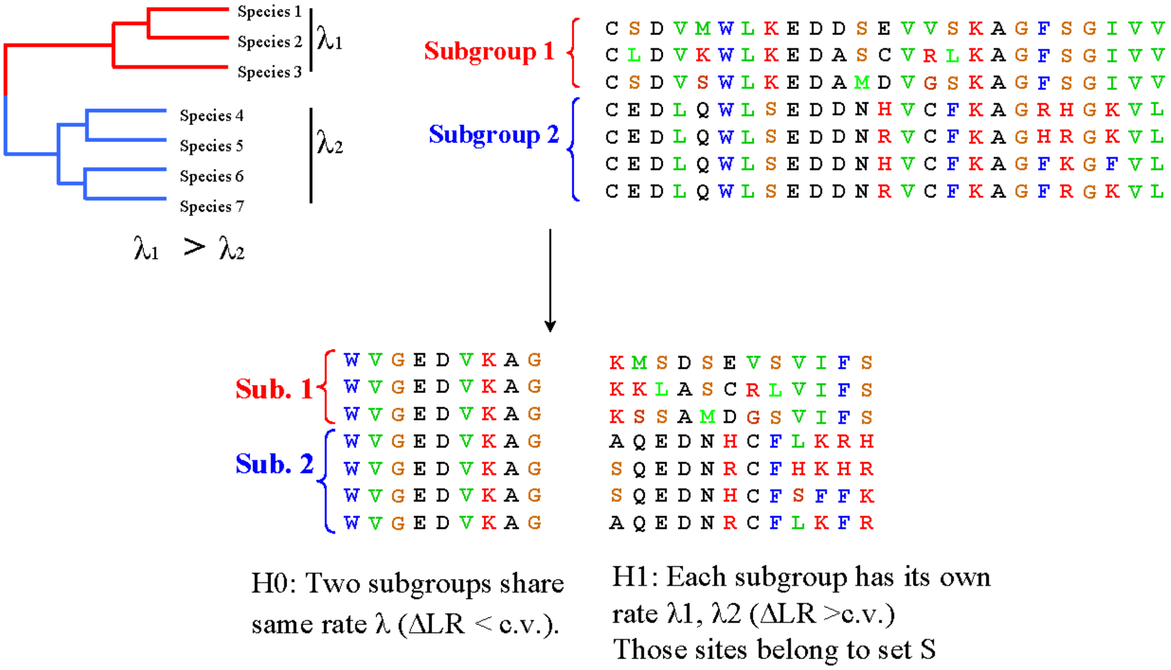

Species 1
Species 2
Species 3
λ_1
Species 4
Species 5
Species 6
Species 7
λ_2
$\lambda_1 > \lambda_2$
Subgroup 1
C S D V M W L K E D D S E V V S K A G F S G I V V
C L D V K W L K E D A S C V R L K A G F S G I V V
C S D V S W L K E D A M D V G S K A G F S G I V V
Subgroup 2
C E D L Q W L S E D D N H V C F K A G R H G K V L
C E D L Q W L S E D D N R V C F K A G H R G K V L
C E D L Q W L S E D D N H V C F K A G F K G F V L
C E D L Q W L S E D D N R V C F K A G F R G K V L
Sub. 1
W V G E D V K A G
W V G E D V K A G
W V G E D V K A G
Sub. 2
W V G E D V K A G
W V G E D V K A G
W V G E D V K A G
W V G E D V K A G
K M S D S E V S V I F S
K K L A S C R L V I F S
K S S A M D G S V I F S
A Q E D N H C F L K R H
S Q E D N R C F H K H R
S Q E D N H C F S F F K
A Q E D N R C F L K F R
H0: Two subgroups share same rate λ (ΔLR < c.v.).
H1: Each subgroup has its own rate λ1, λ2 (ΔLR >c.v.)
Those sites belong to set S

Figure 2

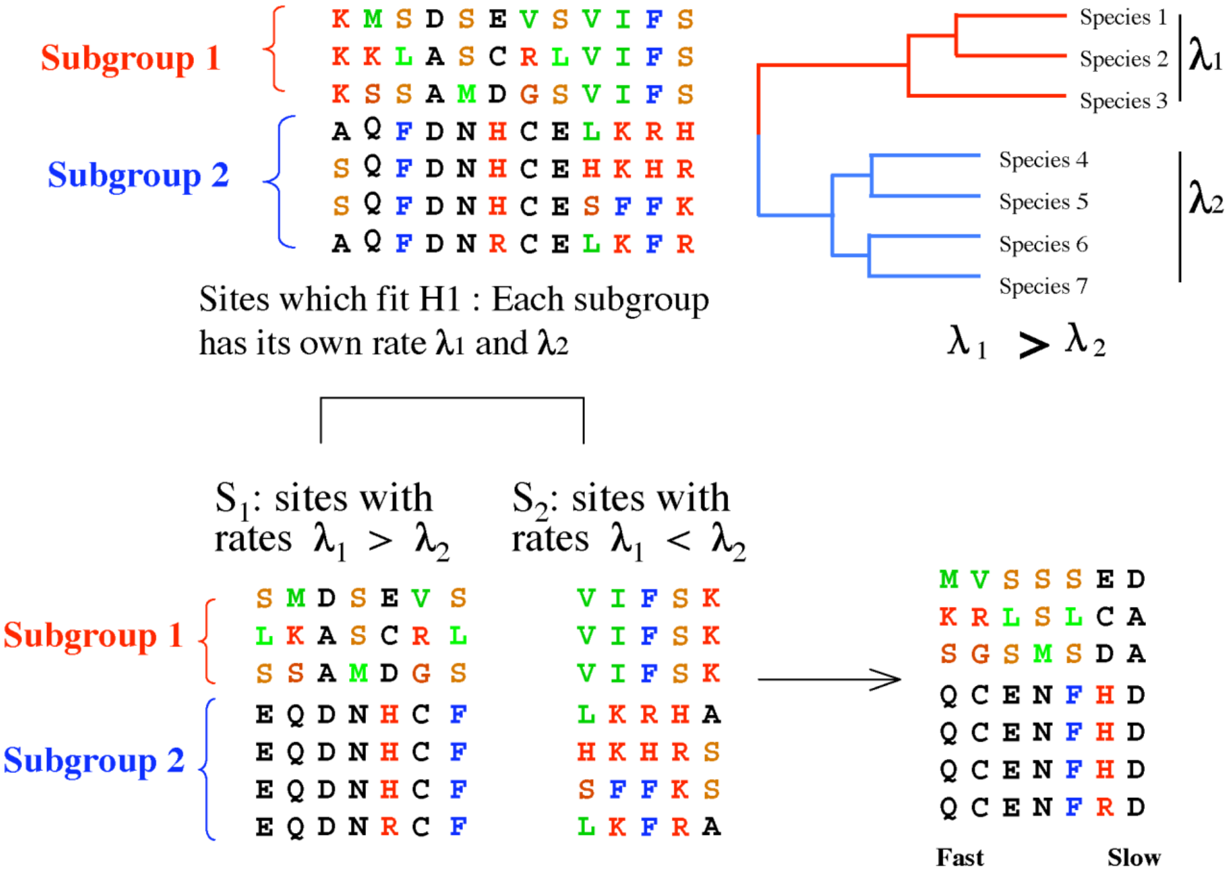

Subgroup 1
K M S D S E V S V I F S
K K L A S C R L V I F S
K S S A M D G S V I F S
Subgroup 2
A Q F D N H C E L K R H
S Q F D N H C E H K H R
S Q F D N H C E S F F K
A Q F D N R C E L K F R
Species 1
Species 2
Species 3
λ_1
Species 4
Species 5
Species 6
Species 7
λ_2
Sites which fit H1 : Each subgroup has its own rate λ_1 and λ_2
$\lambda_1 > \lambda_2$
S_1: sites with rates $\lambda_1 > \lambda_2$
S_2: sites with rates $\lambda_1 < \lambda_2$
Subgroup 1
Subgroup 2
S M D S E V S
L K A S C R L
S S A M D G S
E Q D N H C F
E Q D N H C F
E Q D N H C F
E Q D N R C F
V I F S K
V I F S K
V I F S K
L K R H A
H K H R S
S F F K S
L K F R A
M V S S S E D
K R L S L C A
S G S M S D A
Q C E N F H D
Q C E N F H D
Q C E N F H D
Q C E N F R D
Fast
Slow

Figure 3

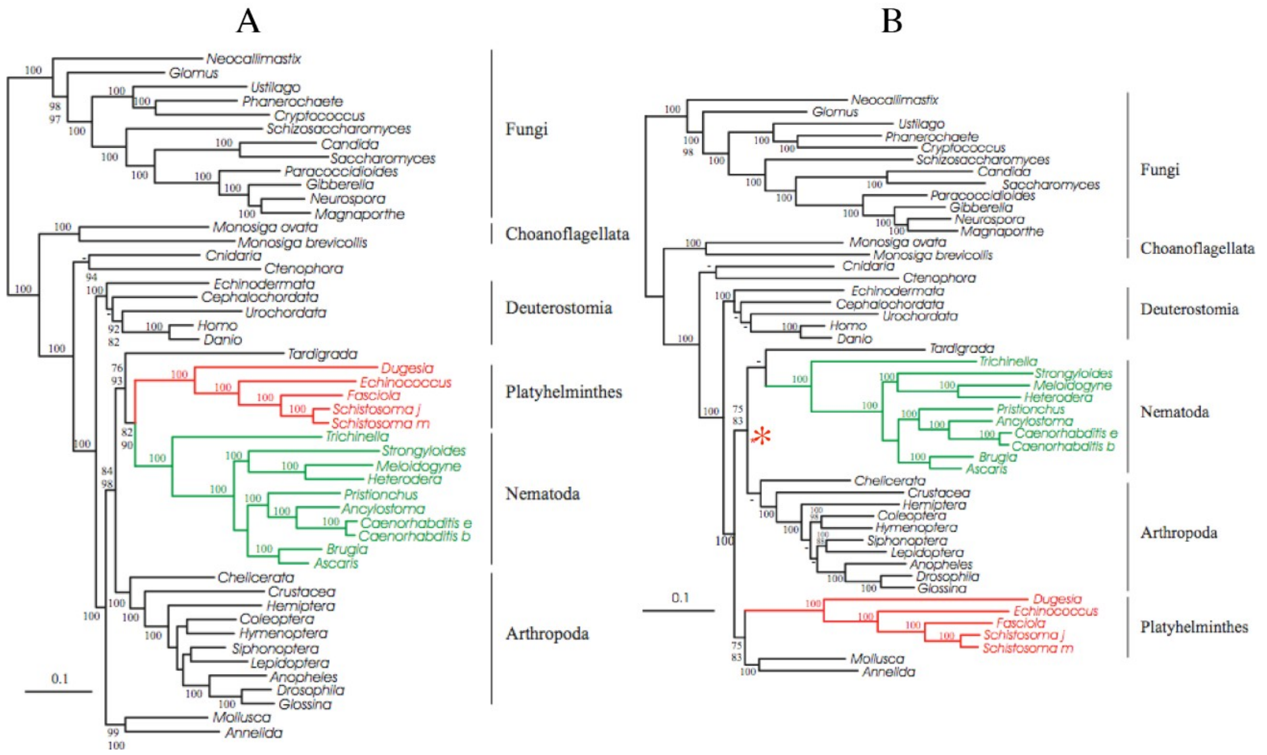
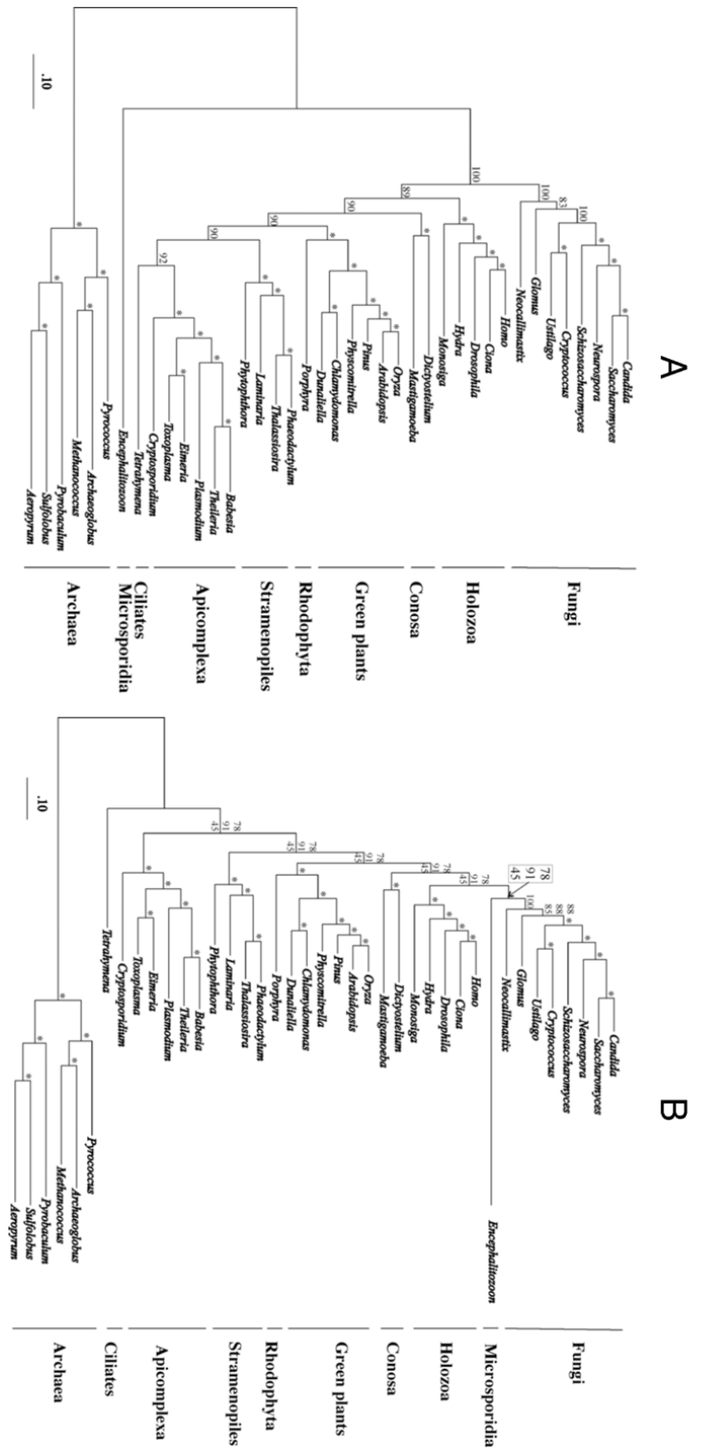

A
B
Neocallimastix
Glomus
Ustilago
Phanerochaete
Cryptococcus
Schizosaccharomyces
Candida
Saccharomyces
Paracoccidioides
Gibberella
Neurospora
Magnaporthe
Monosiga ovata
Monosiga brevicollis
Cnidaria
Ctenophora
Echinodermata
Cephalochordata
Urochordata
Homo
Danio
Tardigrada
Dugesia
Echinococcus
Fasciola
Schistosoma j
Schistosoma m
Trichinella
Strongyloides
Meloidogyne
Heterodera
Pristionchus
Ancylostoma
Caenorhabditis e
Caenorhabditis b
Brugia
Ascaris
Chelicerata
Crustacea
Hemiptera
Coleoptera
Hymenoptera
Siphonoptera
Lepidoptera
Anopheles
Drosophila
Glossina
Mollusca
Annelida
Fungi
Choanoflagellata
Deuterostomia
Platyhelminthes
Nematoda
Arthropoda
0.1
*

Figure 4


A
.10
100
100
83
100
89
90
90
90
92
Candida
Saccharomyces
Neurospora
Schizosaccharomyces
Cryptococcus
Ustilago
Glomus
Neocallimastix
Homo
Ciona
Drosophila
Hydra
Monosiga
Dictyostelium
Mastigamoeba
Oryza
Arabidopsis
Pinus
Physcomitrella
Chlamydomonas
Dunaliella
Porphyra
Phaeodactylum
Thalassiosira
Laminaria
Phytophthora
Babesia
Theileria
Plasmodium
Eimeria
Toxoplasma
Cryptosporidium
Tetrahymena
Encephalitozoon
Pyrococcus
Archaeoglobus
Methanococcus
Pyrobaculum
Sulfolobus
Aeropyrum
Fungi
Holozoa
Conosa
Green plants
Rhodophyta
Stramenopiles
Apicomplexa
Ciliates
Microsporidia
Archaea
B
.10
78
91
45
100
88
88
85
Candida
Saccharomyces
Neurospora
Schizosaccharomyces
Cryptococcus
Ustilago
Glomus
Neocallimastix
Encephalitozoon
Homo
Ciona
Drosophila
Hydra
Monosiga
Dictyostelium
Mastigamoeba
Oryza
Arabidopsis
Pinus
Physcomitrella
Chlamydomonas
Dunaliella
Porphyra
Phaeodactylum
Thalassiosira
Laminaria
Phytophthora
Babesia
Theileria
Plasmodium
Eimeria
Toxoplasma
Cryptosporidium
Tetrahymena
Pyrococcus
Archaeoglobus
Methanococcus
Pyrobaculum
Sulfolobus
Aeropyrum
Fungi
Microsporidia
Holozoa
Conosa
Green plants
Rhodophyta
Stramenopiles
Apicomplexa
Ciliates
Archaea

References

Baldauf, S. L., and J. D. Palmer. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A **90**:11558-11562.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972-977.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A **99**:1414-1419.

Bevan, R. B., D. Bryant, and B. F. Lang. 2007. Accounting for gene rate heterogeneity in phylogenetic inference. Syst Biol **56**:194-205.

Bevan, R. B., B. F. Lang, and D. Bryant. 2005. Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Syst Biol **54**:900-915.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol **16**:817-825.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol **54**:743-757.

Cavalier-Smith, T. 1987. Pp. 339-353 *in* A. D. Rayner, M., C. M. Brasier, and D. Moore, eds. Evolutionary Biology of the Fungi. Symposium of the British Mycological Society. Cambridge Univ. Press, Cambridge, U.K.

Dean, A. M., and G. B. Golding. 2000. Enzyme evolution explained (sort of). Pac Symp Biocomput:6-17.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool **27**:27-33.

Gadagkar, S. R., and S. Kumar. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol **22**:2139-2141.

Gaucher, E. A., and M. M. Miyamoto. 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol Phylogenet Evol **37**:928-931.

Gribaldo, S., D. Casane, P. Lopez, and H. Philippe. 2003. Functional Divergence Prediction from Evolutionary Analysis: A Case Study of Vertebrate Hemoglobin. Mol Biol Evol.

Hartmann, M., and G. B. Golding. 1998. Searching for substitution rate heterogeneity. Mol Phylogenet Evol **9**:64-71.

Keeling, P. J., M. A. Luker, and J. D. Palmer. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol Biol Evol **17**:23-31.

Knudsen, B., and M. M. Miyamoto. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A **98**:14512-14517.

Kolaczkowski, B., and J. W. Thornton. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol **25**:1054-1066.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980-984.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol **19**:1-7.

Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J Mol Evol **49**:496-508.

Pagel, M., and A. Meade. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos Trans R Soc Lond B Biol Sci **363**:3955-3964.

Philippe, H., D. Casane, S. Gribaldo, P. Lopez, and J. Meunier. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life **55**:257-265.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol **22**:1246-1253.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev **8**:616-623.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci **267**:1213-1221.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol **5**:50.

Phillips, M. J., P. A. McLenachan, C. Down, G. C. Gibb, and D. Penny. 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. Syst Biol **55**:122-137.

Pisani, D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst Biol **53**:978-989.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, and P. H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. **56**:389-399.

Roger, A. J., and L. A. Hug. 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. Philos Trans R Soc Lond B Biol Sci **361**:1039-1054.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**:751-754.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol **22**:1161-1164.

Susko, E., Y. Inagaki, C. Field, M. E. Holder, and A. J. Roger. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol Biol Evol **19**:1514-1523.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Chapter 6 Discussion

This thesis contains two different scientific aspects, phylogenomic analyses of real sequence datasets and method development for improved phylogenetic analysis. In the first part, I address several unresolved issues in fungal evolution (for details, see Chapter 1, and (Taylor et al. 2004; James et al. 2006a; Hibbett et al. 2007)) using phylogenomic approaches. These include the construction and comparative analysis of sequence datasets that contain a large number of species and genes, which are derived from both mitochondrial and nuclear genomes (Philippe et al. 2005a). Our results demonstrate that this approach is quite powerful, allowing us to resolve deep divergences within Fungi that have been either controversial and/or without compelling statistical support. The comparisons between mitochondrial and nuclear phylogenies add an additional layer of confidence to our conclusions - in cases where the resulting tree topologies are consistent (Liu et al. 2009b). If different, the results indicate potential phylogenetic artifacts (Ruiz-Trillo et al. 2008; Liu et al. 2009a), which may be analyzed and resolved by applying adequate procedures (e.g., comparison of inferences with methods that differ in their sensitivity towards artifacts (Kolaczkowski and Thornton 2004); removal of fast-evolving or biased sequence sites or genes (Philippe et al. 2000a; Brinkmann et al. 2005)).

In the second part of this thesis, I present a novel method that helps overcoming LBA artifacts (Felsenstein 1978b). It identifies and gradually removes sequence positions that are highly heterotachous (Delsuc, Brinkmann, and Philippe 2005), that contain no or little phylogenetic information, and that are not correctly handled by current phylogenetic inference methods (Zhou et al. 2007). Our proposed method is unique in removing sequences only in fast-evolving species, and I show that it leads to significant improvement of phylogenetic inferences (Chapter 5). In other published procedures, sequence sites are removed column-wise for all species (Philippe et al. 2000a), or complete proteins are eliminated from fast evolving species (Brinkmann et al. 2005). Both leads to significant loss of phylogenetic signal, and less resolution.

In the following chapter I will discuss the major findings and conclusions of this thesis, compare with previously published works, and give an outlook on future directions in this research domain.

Are Zygomycota and Chytridiomycota monophyletic?

Our phylogenomic analyses conclude beyond reasonable doubt that the classic taxon Zygomycota is paraphyletic - as previously suspected (e.g., (Leigh et al. 2003; Taylor et al. 2004; Seif et al. 2005)), and that subdividing this taxon is in order. However, as pointed out in our manuscript (Liu et al. 2009b), dividing and renaming a traditionally as well known taxon as Zygomycota should be based on solid phylogenetic evidence to avoid further renaming, and confusion of the scientific community.

In the given case, I show with high statistical support that two of the new taxonomic definitions (according to (Hibbett et al. 2007)) are either incorrect or not well justified; i.e., the inclusion of *Mortierella* in a phylum Mucoromycotina, and the introduction of a separate phylum for a small group of mycorrhizal fungi known as Glomeromycota, whose phylogenetic position relative to other zygomycete phyla remains unresolved with any of the currently available sequence datasets. The genetics of mycorrhizal fungi is indeed most unusual, with cells containing populations of hundreds of genetically different nuclei per cell (e.g. (Hijri and Sanders 2005; Lee et al. 2008; Liu et al. 2009a)), and a reduced capacity for recombination and accelerated sequence and genome evolution. Not surprisingly, the ongoing *Glomus* genome project has failed at the genome assembly step (Martin et al. 2008). Genome sequencing is unlikely to ever finish unless by processing single, separate nuclei, whose DNA is multiplied by whole-genome amplification – an approach currently under development. Taken together, linking this group of Fungi to its free-living relatives includes genome sequencing of more than just one glomeromycotan species, in the context of untangling the classic phylum Zygomycota into phylogenetically well defined sub-groups.

Similarly, also the classic taxon Chytridiomycota (chytrids) may be paraphyletic. One of its subgroups, the Blastocladiales, associates in phylogenetic analyses either with chytrids (e.g., (Seif et al. 2005) or with zygomycetes ((Van der Auwera and De Wachter 1996; Tanabe, Watanabe, and Sugiyama 2005a; Hoffman et al. 2008; Tambor, Ribichich, and Gomes 2008). Only few genera are known within this group (James et al. 2006a), and genome-size data sets are available for two of its close members, *Allomyces* and *Blastocladiella* (Ribichich, Georg, and Gomes 2006; Liu et al. 2009b). Yet, in the above-mentioned new taxonomy (Hibbett et al. 2007), Blastocladiales are separated from chytrids into a new phylum Blastocladiomycota, despite the unresolved controversy on their phylogenetic position. In our most recent phylogenomic analysis with nuclear sequences (Liu et al. 2009b) Blastocladiales group with chytrids in ML analyses, but not when using the CAT model. With the most recent collection of mitochondrial data, the favored topology appears to be with chytrids, whatever the phylogenetic model used (our unpublished results). Accordingly, a decisive answer to this question still remains to be found. To resolve the questions on chytrid monophyly and the identity of Blastocladiomycota, genome sequencing of distant representatives within this clade (e.g., *Physoderma maydis*; (James et al. 2006a)) and within a broad spectrum of zygomycetes will be important.

Taphrinomycotina are monophyletic

Our phylogenetic analyses with 113 nucleus-encoded proteins conclude with high confidence that Taphrinomycotina are monophyletic, and that they are a sister group of Saccharomycotina plus Pezizomycotina. Application of the AU likelihood test (an alternative to bootstrap analysis; to pass the AU test at the recommended p value of 0.05, an equivalent of about 95% bootstrap support is required for the majority of branches) also confirms the monophyly of Taphrinomycotina, rejecting paraphyletic scenarios with high confidence ($p < 0.01$). Yet, there is currently insufficient data to resolve the relationships within Taphrinomycotina, in part because of missing sequences in some species (due to limited amounts of EST data). In addition, genome-size nuclear or mitochondrial data are not available for *Neolecta irregularis*, a putative member of Taphrinomycotina based only on the analysis of rRNA sequences (Nishida and Sugiyama 1993).

Phylogenetic position of protists related to Fungi

Based on our comprehensive analyses with nuclear and mitochondrial genomic data I show that Nucleariida are the closest sister of Fungi (Liu et al. 2009b). A similarly interesting group of protists are Rozellida. SSU rRNA phylogenies indicate that these intracellular pathogens of chytrids might be even closer to Fungi than Nucleariida (Lara, Moreira, and Lopez-Garcia 2009a). Genome sequences from Rozellida plus Nucleariida will be useful for defining traits of the fungal ancestors, and will be most important in establishing a strong outgroup in phylogenetic analysis of the Fungi. The effect of a close outgroup is both strengthening of resolution and reduction of phylogenetic artifacts, ultimately allowing to settle even the dispute over the relatedness between Microsporidia and Fungi.

The phylogenetic placement of Microsporidia has long been debated, but any of the proposed topologies is without compelling significant support (Thomarat, Vivares, and Gouy 2004; Gill and Fast 2006; Lee et al. 2008; Keeling 2009), which is due to their most elevated evolutionary rates (Katinka et al. 2001). It is indeed possible that the phylogenetic information contained in currently available genome sequences remains insufficient, and that at least one slowly evolving member of Microsporidia has to be found to resolve this question (Akiyoshi et al. 2009; Cornman et al. 2009; Corradi et al. 2009). The combination of primary sequence with other genomic information, like gene order and content, is another strategy to overcome current limitations (Larget et al. 2005; Lavrov and Lang 2005). A recent analysis of gene order conservation in the sex locus and its surroundings suggests that Microsporidia might be related to Mucorales (Lee et al. 2008). Yet, this analysis is based on only few character states, and without the use of statistics. Thus, the development of a statistically valid evolutionary model for these gene order data will be urgently required, to allow their future integration with sequence data.

Advantages and limitations of mitochondrial phylogeny

Due to its functional significance, mitochondrial genes are well conserved across eukaryotes, and are widely used to resolve (sometimes even deep) phylogenies (e.g., (Lang, Gray, and Burger 1999; Lang et al. 2002a; Ballard and Rand 2005)). Furthermore, the comparisons between phylogenies from mitochondrial and nuclear data provide either a confirmation (when both agree), or a valuable indicator of phylogenetic artifacts (when inconsistencies occur). For example, in the case of Nuclearia, our analyses of mitochondrion-encoded genes is consistent with those of nuclear data, even if the statistical support of the mitochondrial phylogeny is moderate due to the small size of dataset ((Liu et al. 2009b), and see below). On the other hand, in the case of Taphrinomycotina, analyses with concatenated mitochondrion-encoded proteins support grouping of *Schizosaccharomyces* and Saccharomycotina, with moderate to strong support (depending on species sampling and on the inference method and phylogenetic model; (Bullerwell et al. 2003; Leigh et al. 2003)). Comparison of topologies with varying taxon sampling, and analysis of datasets from which fast-evolving sites were gradually removed, indicate that the mitochondrial topology is due to an LBA artifact ((Liu et al. 2009a)). Only the slowest-evolving sequence positions in the mitochondrion-encoded proteins recover the tree topology that agrees with the analysis of nuclear data, and then only with marginal statistical support (due to the small amount of data that remains after site removal). To address this issue, either *Schizosaccharomyces* and/or Saccharomycotina species would have to found, which have not undergone such dramatic evolutionary rate acceleration. At present, promising candidate species are unknown.

Another strategy for resolving phylogenies with mitochondrial data is the addition of nucleus-encoded proteins of clearly mitochondrial origin (Cotter et al. 2004; Williams and Keeling 2005; Catalano et al. 2006). For this, nuclear genome sequencing would be required, thus opening the possibility of comparing the consistence of large nuclear *versus* mitochondrial phylogenomic analyses. The difficulty with analyzing a mixed dataset of mtDNA- plus nucleus-encoded mitochondrial genes is in differences of their evolutionary models (evolutionary rate and compositional heterogeneity between genomes), which could be addressed by partitioning of the dataset and the use of separate models (Bapteste et al. 2002; Nylander et al. 2004).

Removing HH sites improves the accuracy of phylogenetic inference

Data removal is universally (implicitly) applied in phylogenetic analyses, by elimination of sequence positions from multiple alignments that are not aligned reliably and contain little if any phylogenetic signal (e.g., either manual removal or by using Gblocks; (Castresana

2000)). This practice is justified because current algorithms for multiple sequence alignment tend to introduce errors, by optimizing alignments in little conserved regions based on sequence bias. The problem is aggravated in very fast evolving species (e.g., Microsporidia), whose sequences are mutationally saturated thus phylogenetically less tractable. In this case, including such species in the dataset may even result in a decrease of quality for the overall alignment. In other words, current technology and models are not well suited to deal with fast evolving sequence sites; inclusion of such sites will introduce potential phylogenetic artifacts. Thus, identification and removal of such positions is expected to improve phylogenetic inference. Previous methods either remove complete fast-evolving proteins of fast-evolving species (Brinkmann et al. 2005), or fast-evolving sequence positions across all species (Philippe et al. 2000a). In both instances, a significant amount of valuable sequence information is discarded, reducing the overall phylogenetic signal to a level that compelling statistical support of the results may no longer be attained. Our approach is different in filtering out accelerated sequence positions only in fast-evolving species that are likely affected by LBA, thus keeping as much (confidently aligned) sequence information as possible. Our studies confirm that this method is superior in improving the accuracy of inferences, but in cases such as the *Schizosaccharomyces*/ Saccharomycotina mitochondrial phylogeny discussed above, insufficient sequence data is left even with our sequence filtering approach.

Reference

Akiyoshi, D. E., H. G. Morrison, S. Lei, X. Feng, Q. Zhang, N. Corradi, H. Mayanja, J. K. Tumwine, P. J. Keeling, L. M. Weiss, and S. Tzipori. 2009. Genomic survey of the non-cultivatable opportunistic human pathogen, Enterocytozoon bieneusi. PLoS Pathog **5**:e1000261.

Ballard, J. W. O., and D. M. Rand. 2005. THE POPULATION BIOLOGY OF MITOCHONDRIAL DNA AND ITS PHYLOGENETIC IMPLICATIONS. Annual Review of Ecology, Evolution, and Systematics **36**:621-642.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A **99**:1414-1419.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol **54**:743-757.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res **31**:1614-1623.

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

Catalano, D., F. Licciulli, A. Turi, G. Grillo, C. Saccone, and D. D'Elia. 2006. MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. BMC Bioinformatics **7**:36.

Cornman, R. S., Y. P. Chen, M. C. Schatz, C. Street, Y. Zhao, B. Desany, M. Egholm, S. Hutchison, J. S. Pettis, W. I. Lipkin, and J. D. Evans. 2009. Genomic analyses of the microsporidian Nosema ceranae, an emergent pathogen of honey bees. PLoS Pathog **5**:e1000466.

Corradi, N., K. L. Haag, J. F. Pombert, D. Ebert, and P. J. Keeling. 2009. Draft genome sequence of the Daphnia pathogen Octosporea bayeri: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. Genome Biol **10**:R106.

Cotter, D., P. Guda, E. Fahy, and S. Subramaniam. 2004. MitoProteome: mitochondrial protein sequence database and annotation system. Nucleic Acids Res **32**:D463-467.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.

Gill, E. E., and N. M. Fast. 2006. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. Gene **375**:103-109.

Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res **111**:509-547.

Hijri, M., and I. R. Sanders. 2005. Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. Nature **433**:160-163.

Hoffman, Y., C. Aflalo, A. Zarka, J. Gutman, T. Y. James, and S. Boussiba. 2008. Isolation and characterization of a novel chytrid species (phylum Blastocladiomycota), parasitic on the green alga Haematococcus. Mycol Res **112**:70-81.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E.

Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature **443**:818-822.

Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature **414**:450-453.

Keeling, P. 2009. Five questions about microsporidia. PLoS Pathog **5**:e1000489.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980-984.

Lang, B. F., M. W. Gray, and G. Burger. 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet **33**:351-397.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr Biol **12**:1773-1778.

Lara, E., D. Moreira, and P. Lopez-Garcia. 2009. The Environmental Clade LKM11 and Rozella Form the Deepest Branching Clade of Fungi. Protist.

Larget, B., D. L. Simon, J. B. Kadane, and D. Sweet. 2005. A bayesian analysis of metazoan mitochondrial genome arrangements. Mol Biol Evol **22**:486-495.

Lavrov, D. V., and B. F. Lang. 2005. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements. Syst Biol **54**:651-659.

Lee, S. C., N. Corradi, E. J. Byrnes, 3rd, S. Torres-Martinez, F. S. Dietrich, P. J. Keeling, and J. Heitman. 2008. Microsporidia evolved from ancestral sexual fungi. Curr Biol **18**:1675-1679.

Leigh, J., E. Seif, N. Rodriguez, Y. Jacob, and B. F. Lang. 2003. Fungal evolution meets fungal genomics. Pp. 145-161 *in* D. K. Arora, ed. Handbook of Fungal Biotechnology. Marcel Dekker Inc., New York.

Liu, Y., J. W. Leigh, H. Brinkmann, M. T. Cushion, N. Rodriguez-Ezpeleta, H. Philippe, and B. F. Lang. 2009a. Phylogenomic analyses support the monophyly of Taphrinomycotina, including Schizosaccharomyces fission yeasts. Mol Biol Evol **26**:27-34.

Liu, Y., E. Steenkamp, H. Brinkmann, L. Forget, H. Philippe, and B. Lang. 2009b. Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. BMC Evolutionary Biology **in press**.

Martin, F., V. Gianinazzi-Pearson, M. Hijri, P. Lammers, N. Requena, I. R. Sanders, Y. Shachar-Hill, H. Shapiro, G. A. Tuskan, and J. P. Young. 2008. The long hard road to a completed Glomus intraradices genome. New Phytol **180**:747-750.

Nishida, H., and J. Sugiyama. 1993. Phylogenetic relationships among *Taphrina, Saitoella*, and other higher fungi. Mol Biol Evol **10**:431-436.

Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol **53**:47-67.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. PHYLOGENOMICS. Annual Review of Ecology, Evolution, and Systematics **36**:541-562.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci **267**:1213-1221.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005b. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol **5**:50.

Ribichich, K. F., R. C. Georg, and S. L. Gomes. 2006. Comparative EST analysis provides insights into the basal aquatic fungus Blastocladiella emersonii. BMC Genomics **7**:177.

Ruiz-Trillo, I., A. J. Roger, G. Burger, M. W. Gray, and B. F. Lang. 2008. A phylogenomic investigation into the origin of metazoa. Mol Biol Evol **25**:664-672.

Seif, E., J. Leigh, Y. Liu, I. Roewer, L. Forget, and B. F. Lang. 2005. Comparative mitochondrial genomics in zygomycetes: bacteria-like RNase P RNAs, mobile elements and a close source of the group I intron invasion in angiosperms. Nucleic Acids Res **33**:734-744.

Tambor, J. H., K. F. Ribichich, and S. L. Gomes. 2008. The mitochondrial view of Blastocladiella emersonii. Gene **424**:33-39.

Tanabe, Y., M. Watanabe, and J. Sugiyama. 2005. Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): Insights from molecular phylogenetics. The Journal of general and applied microbiology **51**:9.

Taylor, J., J. Spatafora, K. O'Donnell, F. Lutzoni, T. James, D. Hibbett, D. Geiser, T. Bruns, and M. Blackwell. 2004. The Fungi. Pp. 171-194 *in* M. J. D. Joel Cracraft, ed. Assembling the Tree of Life. Oxford University Press, New York.

Thomarat, F., C. P. Vivares, and M. Gouy. 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 Auwera, G., and R. De Wachter. 1996. Large-subunit rRNA sequence of the chytridiomycete Blastocladiella emersonii, and implications for the evolution of zoosporic fungi. J Mol Evol **43**:476-483.

Williams, B. A., and P. J. Keeling. 2005. Microsporidian mitochondrial proteins: expression in Antonospora locustae spores and identification of genes coding for two further proteins. J Eukaryot Microbiol **52**:271-276.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol **7**:206.

Chapter 7 Conclusions and future directions

This study has addressed three unresolved phylogenetic questions related to the evolution of fungi and their close relatives, and proposes a new method for overcoming LBA artifacts. The following briefly summarizes the results, and provides an outlook on future directions in these areas of research.

By employing methods that are able to reduce the effect of model violations, (e.g., S-F), I conclude for the first time that the analysis of the mitochondrial dataset is plagued by a strong LBA artifact, leading to the paraphyly of Taphrinomycotina (a clearly monophyletic group with most nuclear sequence datasets). My study resolves this long-standing controversy (Liu et al, 2009a). On the contrary, using nucleariids as outgroup, with a similar nuclear phylogenomic dataset, the traditional Zygomycota become paraphyletic, with significant support (Liu et al, 2009b). This finding is compatible with previous studies that have proposed zygomycete paraphyly (Leigh et al. 2004; Liu, Hodson, and Hall 2006; James et al. 2006), although never at such a high confidence level. Finally, from a more methodological standpoint, inferences of any given dataset may be incorrect due to phylogenetic artifacts, in particular due to different evolutionary rates. My new method that filters HH sites from targeted species will help to identify and reduce potential artifacts (Chapter 5). The proposed procedure makes few assumptions, has a strong statistical basis, and may efficiently overcome LBA as demonstrated.

My studies on zygomycetes not only confirm their paraphyly, but open a discussion on the number of paraphyletic zygomycete lineages, and the proper association of species to them. Other phylogenetic questions, like the branching order of chytrid subgroups remains unresolved (Hibbett et al., 2007), as is the positioning of the very fast evolving Microsporidia either within Fungi (close to zygomycetes (Keeling 2003; Lee et al. 2008)) or as a fungal sistergroup (James et al. 2006a; Liu, Hodson, and Hall 2006). For addressing any of these questions, missing genome data and insufficient taxon sampling have been major obstacles (Philippe et al., 2005). Complete (or almost complete) genome data would have been preferable over EST data. With the rapid increase of fast, high volume genome sequencing, investment into software development that aims at improving automated genome assembly and gene annotation will become imperative. Therefore, although the issue of missing data will be resolved with new sequencing technologies, challenges with mastering the data flood will increase, further amplified by a lower accuracy of genome assemblies and gene annotations in some instances (due to inferior length and quality of the underlying sequence readings).

Another concern is at the level of phylogenetic inference methodology itself. The increase of sequence information leads to both, opportunity to better resolve phylogenies but also a stronger impact of phylogenetic artifacts. Previously unresolved phylogenies might appear to be resolved and statistically well supported, but are incorrect. I and others show that data filtering from the target groups may identify and potentially overcome artifacts (see Chapter 5, page 161; (Inagaki et al. 2004; Brinkmann et al. 2005)). In combination with new inference algorithms, the use of improved evolutionary models (e.g., PhyloBayes, CAT model; (Lartillot and Philippe 2004; Lartillot and Philippe 2008)), the reduction of missing data, and a much improved species sampling, even questions as difficult as the microsporidian phylogenetic position may then be resolved.

Reference

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. **54**:743-757.

Hibbett, D. S., M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lucking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Koljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schussler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao, and N. Zhang. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res **111**:509-547.

Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol Biol Evol. **21**:1340-9.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G. H. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schussler, J. E. Longcore, K. O'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister,

D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, and R. Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature **443**:818-822.

Keeling, P. 2009. Five questions about microsporidia. PLoS Pathog **5**:e1000489.

Lartillot, N., H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095-109.

Lartillot, N., H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci.363:1463-72.

Lee, S. C., N. Corradi, E. J. Byrnes, 3rd, S. Torres-Martinez, F. S. Dietrich, P. J. Keeling, and J. Heitman. 2008. Microsporidia evolved from ancestral sexual fungi. Curr Biol. **18**:1675-1679.

Leigh, J., E. Seif, N. Rodriguez, Y. Jacob, and B. F. Lang. 2003. Fungal evolution meets fungal genomics. Pp. 145-161 in D. K. Arora, ed. Handbook of Fungal Biotechnology. Marcel Dekker Inc., New York.

Liu Y, J. W. Leigh, H. Brinkmann, M. T. Cushion, N. Rodriguez-Ezpeleta, H. Philippe, B. F. Lang. 2009. Phylogenomic analyses support the monophyly of Taphrinomycotina, including Schizosaccharomyces fission yeasts. Mol Biol Evol. **26**:27-34.

Liu, Y, E. T. Steenkamp, H. Brinkmann, L. Forge, H. Philippe, B. F. Lang, 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. BMC Evol Biol. **25**;9:272.

Philippe,H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. PHYLOGENOMICS. Annual Review of Ecology, Evolution, and Systematics. **36**: 541-562

Appendix Contribution to Chapter 2-5

Chapter two:

Constructed the sequence alignment containing 13 mitochondrial proteins, filtered sections that cannot be aligned reliably, and conducted phylogenetic analyses.

Chapter three and four:

For the nuclear dataset, added sequences from fungal species to a previous alignment, identified and removed paralogs, conducted phylogenetic analyses;
For the mitochondrial dataset, constructed the alignment, conducted phylogenetic analyses and designed the methods to detect LBA for mitochondrial dataset.

Chapter five:

Take full responsibility for the whole work: design, implementation of the method, and data analyses.