Towards a reconstruction of the microsporidian last common ancestor gene set

Dissertation zur Erlangung

des Doktorgrades der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Ngoc Vinh Tran

aus Lam Dong, Vietnam

Frankfurt (2018)

(D30)

vom Fachbereich Biowissenschaften der

Johann Wolfgang Goethe-Universität als Dissertation angekommen.

Dekan:	Prof. Dr. Sven Klimpel Institut für Ökologie, Evolution und Diversität Johann Wolfgang Goethe-Universität D-60438 Frankfurt am Main
Gutachter:	Prof. Dr. Ingo Ebersberger AK für Angewandte Bioinformatik Institut für Zellbiologie und Neurowissenschaft Johann Wolfgang Goethe-Universität D-60438 Frankfurt am Main
	Prof. Dr. Enrico Schleiff Abteilung Molekulare Zellbiologie der Pflanzen Institut für Molekulare Biowissenschaften Johann Wolfgang Goethe-Universität D-60438 Frankfurt am Main

Datum der Disputation: ______________________________

Summary

Microsporidia are a group of parasites that infect a wide range of species, many of which play important roles in agriculture and human disease. At least 14 microsporidian species have been confirmed to cause potentially life-threatening infectious diseases in both immunocompromised and immunocompetent humans. Approximately 1,400 species of microsporidia have been described. Depending on their host and habitat they are classified into three groups, the aquasporidia, the terresporidia and the marinosporidia.

Microsporidia were originally classified as fungi by Naegeli (1857). However, their lack of typical eukaryotic components – such as mitochondria, Golgi bodies or peroxisomes – suggested to place the microsporidia together with other amitochondriate protists within the Archezoa kingdom. This "microsporidia-early" hypothesis was further supported by molecular phylogenies inferred from individual genes. Despite this evidence, the placement of microsporidia as an early branching eukaryote remained a topic for debate. The phylogeny of microsporidia is prone to suffer from biases in their reconstruction. The high evolutionary rate of microsporidian proteins tends to place these proteins together with other fast evolving lineages, a phenomenon known as long-branch attraction. In 1996, the first molecular phylogenetic studies placed the microsporidia inside the fungi. Subsequently, several further studies located the microsporidia at different positions inside the fungal clade. Since then, microsporidia have been considered as members of the Ascomycota, Zygomycota, Cryptomycota, or as a sister group to the Ascomycota and Basidiomycota, or even as the sister group of all fungi.

The difficulties in determining the evolutionary origin of microsporidia are not only caused by their lack of several cellular components but also by their

reduced genomes and metabolism. Being obligate intracellular parasites, microsporidia successfully reduced their genome sizes, down to the range of bacteria. As the smallest eukaryotic genome described so far, the genome of *Encephalitozoon intestinalis* is just 2.3 Mbp, about half the size of the one of *Escherichia coli*. Due to their low number of protein coding genes (less than 4,000), microsporidia are thought to retain only genes essential for their survival and development. Furthermore, several key metabolic pathways are missing in the microsporidia, such as the citric acid cycle, oxidative phosphorylation, or the *de novo* biosynthesis of nucleotides. As a result they are in an obligatory dependence on many primary metabolites from the hosts. However, the presence of hsp70 protein suggests a more complex genome of the microsporidian ancestor. Consequently, the small microsporidian genomes and the reduced metabolism would be consequences of a secondary loss process that molded the contemporary microsporidia from a functionally more complex ancestral species. However, it remains unclear whether the last common ancestor (LCA) of the microsporidia was already reduced, or whether the genome compaction was lineage-specific and started from a more complex LCA.

We investigated the evolutionary history of the contemporary microsporidia through the reconstruction and analysis of their LCA. As a first step in our analysis, we have developed and implemented a software facilitating an intuitive data analysis of the large presence absence-patterns resulting from the tracing of microsporidian proteins in gene sets of many different species. These so called phylogenetic profiles can now be dynamically visualized and explored with *PhyloProfile*. The software allows the integration of other additional information layers into the phylogenetic profile, such as the similarity of feature architecture (FAS) between the protein under study and its orthologs. The FAS score can be displayed along the presence-absence pattern, which can help to identify orthologs that have likely diverged in function. PhyloProfile closes the methodological gap that existed between tools to generate large phylogenetic profiles to delineate the evolutionary history and the contemporary distribution of large – and ultimately complete – gene sets, and the more function-oriented analysis of individual protein. In the next step we tackled the problem of how to transfer functional annotation from one protein to another. We have developed *HamFAS* that integrates a targeted ortholog search based on the HaMStR algorithm with a weighted assessment of feature architecture similarities (FAS) between orthologs. In brief, for a seed protein we identify orthologs in reference species in which proteins have been functionally annotated based on manually curated assignments to KEGG Ortholog (KO) groups. The FAS scores between the orthologs and seed proteins are calculated. Subsequently, we compute pairwise FAS scores for all reference proteins within a KO group. A group's mean FAS score serves then as cutoff that must be exceeded to warrant transfer of its KO identifier to the seed. A benchmark using a manually curated yeast protein set showed that HamFAS yields the best precision (98.5%) when compared with two state-of-the-art annotation tools, KAAS and BlastKOALA. Furthermore, HamFAS achieves a higher sensitivity. On average HamFAS annotates almost 50% more proteins than KAAS or BlastKOALA.

With this extended bioinformatics toolbox at hand, we aimed at reconstructing the evolutionary history of the microsporidia. We generated a robust phylogeny of microsporidia using a phylogenomics approach. As a data basis, we identified a set of microsporidian proteins encoded by 80 core genes with one-to-one orthologs. A maximum likelihood analysis of this data with 48 fungi and additionally in 13 species from more distantly related such as animals and plants combined in a supermatrix strongly supported the hypothesis that microsporidia form the sister group of the fungi. We confirmed that the data explains this microsporidia-fungi relationship significantly better than any other of the previously proposed phylogenetic hypotheses.

On the basis of this phylogeny, and of the phylogenetic profiles of microsporidian proteins, we then focused on reconstructing the dynamics microsporidian genome evolution. Between 2% of the proteins in the compact microsporidia *Encephalitozoon intestinalis* and up to 49% of the proteins of *Edhazardia aedis* are private for individual microsporidian species. A comparison of the sequence characteristics of these proteins to that of proteins with orthologs in other microsporidian species revealed individual differences. Yet, without further evidences it remains unclear whether these private genes are indeed lineage-specific innovations contributing to the adaptation of each microsporidium to its host, or whether these are artifacts introduced in the process of gene annotation. A total of 14,410 microsporidian proteins could then be grouped into 1605 orthologous groups that can be traced back to the last common ancestor of the microsporidia (LCA set). We found that 94% of the microsporidian LCA proteins could be tracked back to the last eukaryotic common ancestor. The high evolutionary age of these proteins, together with the resistance against gene loss in the microsporidia suggests that the corresponding functions are essential for eukaryotic life. Further 3% of the LCA proteins could be dated to the common ancestor microsporidia share with the fungi. Only 3% of the LCA proteins appear as microsporidia specific inventions. These proteins are potentially of importance for the evolutionary of the obligate parasitic lifestyle nowadays shared by all microsporidia.

The functional annotation and metabolic pathway analysis of the microsporidian LCA protein set gave us more insight into the adaptation of the microsporidia to their parasitic lifestyle and the origin of the microsporidian genome reduction. The presence of E1 and E3 components of the pyruvate dehydrogenase complex and the mitochondrial hsp70 protein support an ancestral presence of mitochondria in the ancestral microsporidia. In addition, several ancient proteins that complement gapped metabolic pathways were found in the microsporidian LCA. They suggested a more complex genome and metabolism in the LCA. However, our reconstruction of the metabolic network of the microsporidian LCA still lacks many main pathways. For example, the TCA cycle for effective energy production, and key enzymes that are required for *in vivo* synthesis of critical metabolites like purines and pyrimidines appear absent. We therefore find that the parasitic lifestyle and the genome reduction already occurred in the microsporidian LCA. This ancestral state was followed by further losses and gains during the evolution of each individual microsporidian lineage.

Zusammenfassung

Mikrosporidien sind sporenbildende Parasiten, die verschiedene Organismen infizieren. Die Mikrosporidiose beeinträchtigt nicht nur die Agrarökonomie, sondern führt auch zu humanmedizinischen Krankheitsbildern. Es sind mindestens 14 Mikrosporidien bekannt, welche potenziell lebensbedrohliche infektiöse Krankheiten sowohl bei immunkompromittierten als auch bei immunkompetenten Menschen verursachen. Etwa 1.400 Mikrosporidien wurden bislang beschrieben. Nach ihren Wirten und Habitaten können sie in drei Gruppen eingeteilt werden, die Aquasporidien, Terresporidien und Marinosporidien.

Mikrosporidien wurden erstmals 1857 von Naegeli als Pilz klassifiziert. Wegen ihres Mangels vieler typischen eukaryotischen Komponenten – wie Mitochondrien, Golgi-Apparat oder Peroxisomen – wurden die Mikrosporidien später allerdings zusammen mit anderen amitochondrischen Protisten innerhalb des Archezoa-Reiches gruppiert. Diese "Microsporidia-early" Hypothese wurde darüber hinaus durch einzelgenbasierte molekulare Phylogenien unterstützt. Trotz dieser Evidenzen wurde die phylogenetische Platzierung der Mikrosporidien in Frage gestellt. Die Phylogenie von Mikrosporidien werden durch die Rekonstruktionsartefakte verzerrt. Durch die hohe Evolutionsrate der mikrosporidischen Proteine gruppieren sie häufig zusammen mit anderen schnell evolvierenden Proteinen (long-branch-attraction). Im Jahr 1996 wurde die Verwandtschaft zwischen Pilzen und Mikrosporidien erstmals durch molekulare phylogenetische Studien unterstützt. Verschiedene Studien, basierend auf einzelnen und mehreren Genen, positionieren die Mikrosporidien unterschiedlich in dem Stammbaum der Pilze. Dabei werden die Mikrosporidien entweder innerhalb der Ascomycota, Zygomycota oder Crytomycota positioniert, oder als

Schwestergruppe der Ascomycota und Basidiomycota, oder auch aller Pilze.

So bleibt die exakte Position der Mikrosporidien im Speziesbaum der Pilze immer noch ungelöst. Dabei wird die Bestimmung des Ursprungs der Mikrosporidien durch deren reduzierten Genome weiter erschwert. Als obligate intrazelluläre Parasiten verminderten die Mikrosporidien ihre Genome soweit das deren Größe im Bereich bakterieller Genome liegt. Das kleinste beschriebene eukaryotische Genom von *Encephalitozoon intestinalis* ist mit 2,3 Mbp etwa halb so groß wie das von *Escherichia coli*. Die geringe Anzahl von protein-kodierenden Genen (weniger als 4.000) deutet darauf hin, dass die Genome der Mikrosporidien nur Gene enthalten die für ihr Überleben und ihre Entwicklung essentiell sind. Darüber hinaus fehlen den Mikrosporidien mehrere Stoffwechselwege, wie der Zitronensäurezyklus, die oxidative Phosphorylierung oder die *de novo* Biosynthese von Nukleotiden. Dies führt zu einer obligaten Abhängigkeit vom Wirt für viele primäre Metabolite. Das Vorhandensein des hsp70-Proteins setzt jedoch ein komplexeres Genom des mikrosporidischen Vorfahren voraus. Folglich wären die kleinen mikrosporidischen Genome und der reduzierte Metabolismus die Konsequenzen eines sekundären Verlustprozesses, der die heutigen Mikrosporidien aus einer funktionell komplexeren angestammten Spezies geformt hat. Es bleibt jedoch unklar, ob der mikrosporidische letzte gemeinsame Vorfahr (LCA) bereits reduziert wurde oder ob die Genomkomprimierung linienspezifisch war und von einem komplexeren LCA ausging.

Wir untersuchten daher die Entwicklungsgeschichte der heutigen Mikrosporidien durch die Rekonstruktion und Untersuchung ihres LCAs. Im ersten Schritt unserer Analyse haben wir ein Programm für eine intuitive Datenanalyse eines großen An- und Abwesenheitsmusters entwickelt. Das Muster ist das Ergebnis einer evolutionären Zurückverfolgung von mikrosporidischen Proteinen in Gensets verschiedener Spezies. Mit *PhyloProfile* können diese sogenannten phylogenetischen Profile nun dynamisch visualisiert und untersucht werden. Außerdem erlaubt das Programm die Einbindung von zusätzlichen Informationsebenen zum Profil, wie beispielsweise der Feature Architektur Ähnlichkeit (FAS) zwischen dem untersuchten Protein und seinen Orthologen. Der FAS Wert kann neben dem An- und Abwesenheitsmuster angezeigt werden, was dabei helfen kann Orthologe zu identifizieren, deren Funktion wahrscheinlich divergiert ist. PhyloProfile schließt damit eine methodologische Lücke um die evolutionäre Geschichte und die gegenwärtige Verteilung großer – auch vollständiger – Gensets und die funktionalere Analyse einzelner Proteine zu beschreiben. Im nächsten Schritt haben wir uns mit dem Problem befasst, wie man eine funktionale Annotation von einem Protein zum anderen übertragen kann. Dafür entwickelten wir HamFAS, einen neuen Ansatz der eine gezielte Orthologensuche basierend auf dem HaMStR-Algorithmus mit einer gewichteten Bewertung von Feature Architektur Ähnlichkeiten (FAS) zwischen Orthologen integriert. Für ein Seed-Protein identifizieren wir Orthologe in Referenzspezies, deren Proteine bereits durch eine manuelle Annotation in KEGG-Ortholog (KO)-Gruppen eingeordnet wurden. Zwischen den Orthologen und den Seed-Proteinen werden die FAS-Werte berechnet. Anschließend berechnen wir paarweise FAS-Werte für alle Referenzproteine innerhalb einer KO-Gruppe. Der mittlere FAS-Wert einer Gruppe dient dann als Cutoff, der überschritten werden muss, um die Übertragung seines KO-Identifikation an den Seed zu rechtfertigen. Wir benchmarkten die Performance von HamFAS mit einem manuell kuratierten, KO-annotierten, Hefeprotein-Set. HamFAS erzielte die beste Genauigkeit (98,5%) im Vergleich zu zwei State-of-the-Art Annotationsprogrammen KAAS und BlastKOALA. Darüber hinaus zeigte HamFAS eine höhere

Sensitivität. Hier annotierte HamFAS fast 50% mehr Proteine als KAAS oder BlastKOALA.

Mit diesen beiden Programmen haben wir die Entwicklungsgeschichte der Mikrosporidien rekonstruiert. Wir identifizierten ein evolutionär konserviertes mikrosporidisches Genset, welches aus 80 eins-zu-eins-Orthologen Gruppen besteht. Anschließend erstellten wir eine robuste Phylogenie der Mikrosporidien aus dem Genset zusammen mit den Daten von 48 Pilzen und 13 zusätzlich Spezies von weiter entfernten Verwandten, wie Tieren und Pflanzen. Diese Maximum-Likelihood-Analyse, die in einer Supermatrix kombiniert ist, unterstützte die Hypothese, dass Mikrosporidien die Schwestergruppe der Pilze bilden. Die analysierten Daten erklärten diese Mikrosporidien-Pilz-Verwandtschaft signifikant besser als alle anderen Hypothesen.

Auf der Grundlage dieser Phylogenie und der phylogenetischen Profile mikrosporidischer Proteine rekonstruierten wir die Dynamik der Genomentwicklung. Je nach Mikrosporidium finden wir das zwischen 2% der Proteine in der kompakten Mikrosporidie *Encephalitozoon intestinalis* bis hin zu 49% der Proteine im Fall von *Edhazardia adis* nur in einer Art gefunden werden. Ein Vergleich der Sequenzeigenschaften zwischen diesen Proteinen und den Proteinen, die Orthologe in anderen Spezies haben, zeigte individuelle Unterschiede. Dennoch bleibt es ohne weitere Hinweise ungewiss, ob diese exklusiven Gene tatsächlich abstammungslinienspezifische Gene zur Wirtsanpassung sind oder ob sie Artefakte des Genannotationsprozesses sind. Insgesamt konnten 14,410 mikrosporidische Proteine zu 1605 orthologen Gruppen zusammengefasst werden, die zum LCA der Mikrosporidia (LCA Set) zurückverfolgt werden konnten. Wir finden, dass 94% der Proteinen des mikrosporidischen LCAs auf den letzten gemeinsamen Vorfahren aller Eukaryoten zurückverfolgt. Das hohe evolutionäre Alter dieser Proteine zusammen mit der Resistenz gegen Genverlust in den Mikrosporidien weist darauf hin, dass die entsprechenden Funktionen essentiell für eukaryotisches Leben sind. Nur 3% LCA-Proteine sind spezifisch für Mikrosporidien. Diese Proteine sind potentiell wichtig für die Evolution der mikrosporidischen parasitischen Lebensweise, die von allen Mikrosporidien geteilt wird.

Die funktionelle Annotation und die Analyse der Stoffwechselwege des mikrosporidischen LCAs ermöglichte ein besseres Verständnis der Anpassung von Mikrosporidien an ihre parasitäre Lebensweise und den Ursprung ihrer Genom-Reduktion. Die Anwesenheit von E1, E3-Komponenten des Pyruvat-Dehydrogenase-Komplexes und des mitochondrialen hsp70-Proteins deuten darauf hin, dass die anzestralen Mikrosporidien Mitochondrien besaßen. Zusätzlich wurden mehrere alte Proteine im mikrosporidischen LCA gefunden, die einige Lücken Stoffwechselwege schließen können. Dies deutet auf ein komplexeres Genom und einen aufwändigeren Metabolismus im LCA hin als bislang vermutet. Dem mikrosporidischen LCA fehlen jedoch weiterhin viele primäre Stoffwechselwege, wie der Citratzyklus, oder Schlüsselenzyme, die für die in vivo Synthese von kritischen Metaboliten wie Purinen und Pyrimidinen benötigt werden. Zusammenfassend nehmen wir an, dass die parasitische Lebensweise bereits in der mikrosporidischen LCA vorkam. Die reduzierten Genome sind damit der anzestrale Zustand für die Mikrosporidien, welchem weitere Genverluste und Genzuwächse auf einzelnen mikrosporidischen Linien folgte.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Microsporidia – A clade of emerging pathogens

Microsporidia are a group of obligate intracellular parasites. As of today, approximately 1,400 species have been reported (Dean, Hirt, and Embley 2016). Depending on the host and environment type, microsporidia are classified into three groups, namely the aquasporidia, the terresporidia and the marinosporidia (Vossbrinck, Debrunner-Vossbrinck, and Weiss 2014).

Microsporidia infect a large variety of invertebrate and vertebrate species, such as hornworm, honeybee, mosquitoes, shrimp, farm-raised fish, and even humans (Weiser 1976; Canning 1986; Vossbrinck et al. 1987; Scanlon et al. 2000; Kmmari et al. 2018). They were discovered as pathogens that are responsible for a broad range of diseases, many of which affect species that are of economical relevance. For example, the first microsporidium described, *Nosema bombycis,* has been identified as the causative agent for the silkworm disease (pébrine) (Pasteur 1870), which has seriously affected the silk industry in the mid-nineteenth century (Vivarès and Méténier 2001). Other species from the same genus, *Nosema apis* and *Nosema ceranae*, cause nosemosis disease on the European honeybee *Apis mellifera*, which substantially affected the commercial honey producers in recent years (Neumann and Carreck 2010; Charbonneau et al. 2016). Likewise, the finfish aquaculture is suffering from infections from *Pseudoloma neurophilia,* and from several species from the genus *Glugea* (Ramsay et al. 2009; Ryan and Kohler 2016). The first described mammalian infection was caused by *Nosema cuniculi* in 1922 – the species was renamed in 1923 to *Encephalitozoon cuniculi* (Weiser 1964). This microsporidium infects brain, spinal cords and kidneys of rabbits (Vivarès and Méténier 2001). Eventually, in 1959 it was detected that microsporidia are also capable of infecting humans, which resulted in a substantial increase of attention for microsporidia. Until now, at least 14 species in different genera, including *Enterocytozoon*, *Encephalitozoon*, *Vittaforma*, *Anncalia*, *Tubulinosema* and *Pleistophora,* have been confirmed to be human-infecting microsporidia (Mathis, Weber, and Deplazes 2005; Vossbrinck, Debrunner-Vossbrinck, and Weiss 2014). Among these, the most prevalent pathogens are *Enterocytozoon bieneusi* and *Encephalitozoon intestinalis* (Santín and Fayer 2011; Ramanan and Pritt 2014). The first report that microsporidia represent an opportunistic pathogen came from Desportes et al. (1985). The authors identified *E. bieneusi* as the causative agent for chronic diarrhea in patients suffering from AIDS (Desportes et al. 1985). Since then, intestinal microsporidiosis caused by *E.bieneusi* has been found in other immunocompromised patients, such as organ transplant recipients, but also travelers, children, and elderly are particularly susceptible for infections (Rogelio et al. 2006; Matos, Lobo, and Xiao 2012). *E.bieneusi* appears to display a considerably narrow infection range, where manifestations appear confined to the intestines, as well as the respiratory and biliary tracts. In contrast, *E.intestinalis* together with other human-infecting microsporidia display a wider spectrum. They are additionally involved in diseases related to kidneys, lungs, eyes and other organs (Mathis, Weber, and Deplazes 2005; Ramanan and Pritt 2014). As a consequence, the catalogue of symptoms from microsporidiosis in immunocompromised patients is considerably large, ranging from chronic diarrhea, hepatobiliary and pulmonary illness to more unspecific manifestations such as general abdominal pain or weight loss (Matos, Lobo, and Xiao 2012; Ramanan and Pritt 2014). Notably, microsporidia can also infect immunocompetent people. Here, they can cause acute, self-limiting diarrhea or ocular infections (Mathis, Weber, and Deplazes 2005). Such as, self-limited diarrhea has been reported in approximate 40% of people travelling from the industrialized countries to the developing nations (Rogelio et al. 2006). 17% of HIV-negative elderly patients in the study of

(Lores et al. 2002) suffered from intestinal microsporidiosis caused by *E.bieneusi*. This disease was also detected in up to 22.5% healthy children in Africa and Asia (Bretagne et al. 1993; Mungthin et al. 2001; Mathis, Weber, and Deplazes 2005; Matos, Lobo, and Xiao 2012). Especially, microsporidia can cause asymptomatic infections in both immunocompetent and immunocompromised patients (Matos, Lobo, and Xiao 2012; Stentiford et al. 2016). Such unspecific symptoms render an early detection and the selection of an appropriate treatment hard.

Human microsporidiosis can be transferred from infected human, animals or contaminated water and food through the fecal-oral route (Santín and Fayer 2011; Matos, Lobo, and Xiao 2012). The latter transmission factors are thought to be the most likely path for microsporidia to enter the human body (Didier and Weiss 2011). Many studies have reported the existence of microsporidian pathogens in water and food for human consumption. Such as, the microsporidian species *E. bieneusi*, *E.hellem* and *E. intestinalis*, were detected in municipal wastewaters in Ireland by (Cheng et al. 2011); *E.bieneusi* and *E.intestinalis* were identified in soft fruits, vegetables, and herbs collected from markets in Poland (Jedrzejewski et al. 2007); milk contaminated with *E.bieneusi* was observed in Korea by (Lee 2008); or *E.bieneusi* in cucumbers caused gastrointestinal illness for more than 100 people in Sweden (Decraene et al. 2012). From the summary of (Stentiford et al. 2016), pathogens transmitted via food and water caused many deaths, notably in children under 15 year olds in low-income countries where 40% of the cases were largely due to infectious diseases, and elderly people over 70 in high-income countries, in which 70% of deaths were the result of chronic conditions.

In addition to the direct transmission via food consumption, microsporidia can also transferred to human from the infected insect by bite or sting (Stentiford et al. 2016). For instance, the mosquito pathogen *Anncaliia algerae* infects eyes and musculature (Coyle et al. 2004); the fruit fly microsporidia

Tubulinosema sp. causes infection in tongue of an immunosuppressed patient (Choudhary et al. 2011); or *Trachipleistophora spp.* infects in skeletal muscle and organs of immunodeficient patients (Mathis, Weber, and Deplazes 2005; Jiří et al. 2007). Furthermore, living close to farming animals like chickens, pigs, cows, or contacting with pets also provides a potential risk for the zoonotic transfer of microsporidia into human (Stentiford et al. 2016).

1.2 The evolutionary origin of microsporidia

The precise evolutionary origins of microsporidia have remained in the dark for quite some time. The first attempt to systematically describe microsporidia, and to determine their position in the tree of life, was made by Naegeli (1857). He described the microsporidium *Nosema bombycis* as a yeast-like unicellular fungus. Some years later, they were moved from the fungi into the phylum Sporozoa in the kingdom Chromista (Balbiani 1882). However, since then the microsporidia experienced several rounds of radical taxonomic revisions (Corradi and Keeling 2009), as have been demonstrated by the dashed lines in Figure 1-1.

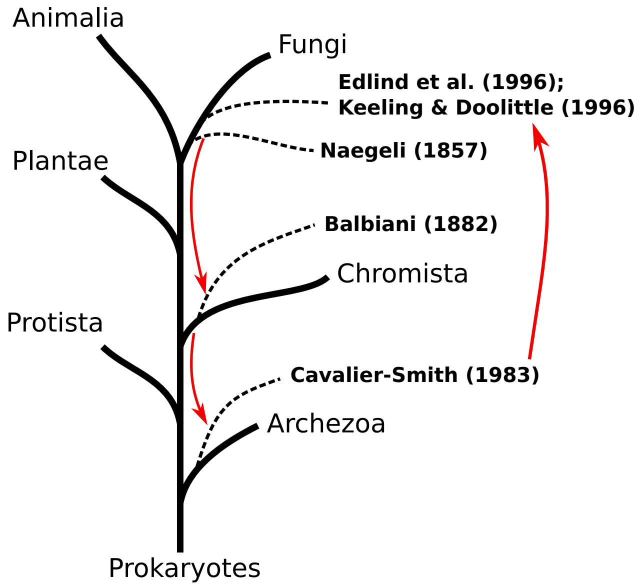

Figure 1-1: The evolutionary relationships of the Eukaryotes according to the Archezoa hypothesis (Cavalier-Smith 1983).

1.2.1 The era of morphology-informed phylogenetic placements

Detailed morphological studies using electron microscopy revealed that microsporidia lack typical eukaryotic components, such as mitochondria, Golgi bodies or peroxisomes (Krieg 1955; Kudo and Daniels 1963; Vavra 1965). As a consequence, they were placed together with other amitochondriate protists into the kingdom Archezoa (Cavalier-Smith 1983). It was assumed that the simple cellular organization of these species reflects the ancestral state of a primordial eukaryote. Consequently, it was hypothesized that the microsporidian lineage separated very early during eukaryotic evolution, and in particular prior to the acquisition of mitochondria.

1.2.2 The era of molecular phylogenies

The second round of taxonomic revision was initiated by the use of molecular markers to reconstruct the evolutionary relationships between sequences and the corresponding species. The first molecular phylogeny providing information about the position of microsporidia in the tree of life was based on the SSU rRNA and LSU rRNAs of the microsporidium *Vairimorpha necatrix* (Vossbrinck et al. 1987). Notably, the resulting tree was in line with the Archezoa hypothesis postulated by Cavalier-Smith (1983). Thus, it seemed to provide a second line of support, next to the morphological characteristics determined by microscopy, that microsporidia are indeed representatives of an ancient lineage that separated early during eukaryotic evolution and diversification. This placement of microsporidia as an early branching eukaryote was then further corroborated by determining the molecular phylogenies of other genes, such as isoleucyl aminoacyl-tRNA synthetase (Brown and Doolittle 1995), elongation factor-1alpha, and elongation factor-2 (Kamaishi et al. 1996).

Despite the seemingly convincing evidences, the "Microsporidia-early" hypothesis remained the matter of a controversial debate (Lee et al. 2008). One of the reasons was, that microsporidia possess the mitochondrial version of the heat shock protein 70 (mtHsp70). The presence of this protein is typically connected with the presence of mitochondria (Germot, Philippe, and Le Guyader 1996). Accordingly, the mtHSP70 of microsporidia was taken as initial hint that this taxonomic group originally possessed mitochondria, and that the absence of this organelle is a consequence of a secondary loss (Germot, Philippe, and Le Guyader 1997; Hirt et al. 1997). As a consequence, one of the main morphological arguments for placing microsporidia at the base of the eukaryotic tree would be void. The more convincing arguments, namely that the placement of microsporidia at the base of the eukaryotic tree might be an artifact rather than reflecting the evolutionary truth, came from a more thorough analysis of possible biases in the reconstruction of phylogenies from molecular data with the algorithms at hand.

Already in the 1970ies, Joe Felsenstein showed that parsimony based approaches to reconstruct phylogenetic trees from molecular data are bound to produce artificial groupings when the homologous proteins on the individual lineages differ substantially in their evolutionary rate (Felsenstein 1978). Importantly, he could show that the algorithms tend to place the fast evolving lineages next to each other, even in such cases where the true phylogeny would place a fast evolving lineage next to a slow evolving one (Figure 1-2).

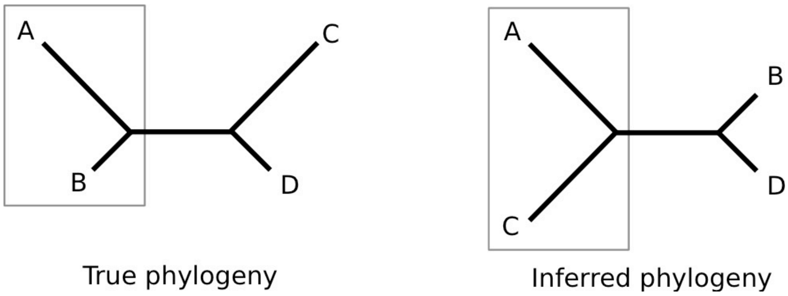

Figure 1-2: Felsenstein's theory about long branch attraction in phylogenetic tree reconstruction. The true phylogeny (left) places the fast evolving species A next to the slow evolving species B. The inferred tree (right), however groups two fast evolving species A and C together.

This artifact was later referred to as *long branch attraction* (LBA) (Bergsten 2005). More importantly, Felsenstein could show that – under the scenario shown in Figure 1-2 – the reconstruction methods become statistically inconsistent. In simple words, the more data is added, the higher gets the confidence in the wrong solution. Although it was initially proposed that maximum likelihood (ML) reconstruction methods do not suffer substantially from statistical inconsistency, it was later shown that this is not always true. In particular, when the model of sequence evolution does not accurately reflect the way how the sequences have evolved – which can be safely assumed to be very often the case – also ML methods can suffer from long branch attraction artifacts (Philippe 2000; Parks and Goldman 2014).

Interestingly, microsporidian sequences bring along the two main characteristics that are required for a phylogeny reconstruction program to end up in artificial placements (Philippe et al. 2005; Kück et al. 2012; Li, Hua, et al. 2014). The evolutionary rate of microsporidian proteins is among the highest in the eukaryotic domain (Thomarat, Vivarès, and Gouy 2004; Corradi and Keeling 2009). Second, there are no species available that could break the long branch that connects microsporidia with the remainder of the eukaryotic diversity. Thus, phylogenetic reconstructions aiming at determining the position of microsporidia in the tree of life are prone to suffer from LBA. As a consequence, the placement of the microsporidia at the base of the eukaryotic tree could be an artifact (Keeling and Fast 2002; Corradi and Keeling 2009).

In 1996 the first molecular phylogenies emerged that revived the original association of microsporidia with fungi. Edlind et al. (1996) and Keeling and Doolittle (1996) analyzed the phylogenies of alpha- and beta-tubulins from several microsporidia species and reported grouping of these sequences with their fungal counterparts, to the exclusion of alpha- and beta-tubulins from other eukaryotes. Thereafter, a number of further molecular studies supported the shared evolutionary descent of fungi and microsporidia. Among the analyzed genes were the mitochondrial Hsp70 (Germot, Philippe, and Le Guyader 1997; Hirt et al. 1997), and the largest subunit of the RNA polymerase II (Hirt et al. 1999).

More recently, the generation of joint phylogenies from more than one gene further supported the fungal association of microsporidia. The analyses included the combined analysis of alpha and beta subunits of pyruvate dehydrogenase E1 (Fast and Keeling 2001), or the DNA-dependent RNA polymerase II largest subunit RPB1 and translation elongation factor I alpha (Tanabe, Watanabe, and Sugiyama 2002).

1.2.3 Do microsporidia fall within or outside the fungal diversity?

With the advent of the phylogenomic era (reviewed in (Ebersberger, von Haeseler, and Schmidt 2007)), the question about the precise position of microsporidia in the tree of life moved again into the focus. The starting point was that, again, analyses using different genes ended up in conflicting hypotheses.

Keeling, Luker, and Palmer (2000) used beta-tubulin to suggest a placement of microsporidia within the fungal diversity. More precisely, they could associate them with either ascomycetes or zygomycetes. Three years later, Keeling (2003) provided another evidence for the zygomycete origin based on the study of alpha and beta tubulin genes. Gene order analyses of a locus allegedly involved in mating was briefly reported as a 'smoking gun' supporting the grouping of microsporidia with zygomycetes (Lee et al. 2008). However, subsequent analyses revealed that the conclusions from Lee et al. (2008) were based on the misinterpretation of the evolutionary relationships of the genes compared between the species (Koestler and Ebersberger 2011). Accordingly, there is currently no evidence that support a close evolutionary relationship of zygomycetes and microsporidia. A combined phylogenetic approach of eight genes including α-tubulin, β-tubulin, the largest subunit of RNA polymerase II (RPB1), the DNA repair helicase RAD25, TATA-box binding protein, a subunit of the E2 ubiquitin conjugating enzyme, and the E1α and β subunits of pyruvate dehydrogenase from Gill and Fast (2006) placed microsporidia as the sister clade to both ascomycetes and basidiomycetes. James et al. (2006), on the other hand, supposed the close relationship between microsporidia and *Rozella allomycis*, a fungus belonging to the Cryptomycota. Nevertheless, none of those studies could reject the placement of microsporidia as the sister group to all fungi using topology tests (James et al. 2006).

Capella-Gutiérrez, Marcet-Houben, and Gabaldón (2012) were among the first to suggest on the basis of a phylogenomic analysis integrating the phylogenetic signal of 53 proteins that microsporidia are the sister group of fungi. As of today, this approach is still among the most comprehensive attempts that aim at decisively solving the issue where microsporidia are placed in the eukaryotic domain. The taxon sampling comprised six microsporidia, twelve species representing six fungal phyla, and an outgroup containing two animals together with two close relatives of the animals, *Monosiga brevicollis* and *Capsaspora owczarzaki.* The resulting tree was then manually rooted with these four outgroup species, by that enforcing the monophyly of fungi and microsporidia. Given the long-standing debate about the evolutionary origin of microsporidia, a more objective way determining the position of microsporidia in the tree would have been desirable. For example, a subset of taxa from outside the opisthokonts (the systematic group uniting animals and fungi) could have been used to confirm that microsporidia group together with the opisthokonts. Subsequently, the position within the opisthokonts could have been easily determined.

To date, after more than 150 years from the report of Naegeli (1857), microsporidia are meanwhile re-classified as fungi (Hibbett et al. 2007) by placing them either within or in the earliest branch of the fungal clade.

1.3 The symbiotic lifestyle of microsporidia

The lack of several cellular components that are characteristic for members of the eukaryotic domain makes microsporidia particularly special eukaryotes. In the course of adaptation to the endoparasitic lifestyle, microsporidia have specialized to an extent that they no longer can exist – apart from the inactive stage (spore) – outside of their host's cell (Garcia 2002). The sporoplasm of the microsporidian spore is transferred into the host cell through its polar tube (Fast and Keeling 2001). The meront, the development state of microsporidian cell, divides and grows inside the host cytoplasm or nuclei until a mature spore is differentiated and exits the host cell to begin a new infection cycle (Scanlon et al. 2000; Vivarès and Méténier 2001; Dean, Hirt, and Embley 2016).

The obligate intracellular parasitic lifestyle of microsporidia is a challenge for studying their physiology. The electropermeabilization through the microsporidian spore wall was unsuccessful (Bohne, Böttcher, and Groß 2011). Moreover, no method for genetic modification in microsporidia exists (Weiss and Becnel 2014). This leaves comparative genome analyses as the method of choice for investigating evolution and function of this enigmatic clade.

1.4 Microsporidia are showcases for the secondary reduction of genomes and the encoded functions

Much of the difficulties in determining the evolutionary origins of microsporidia can be explained by their extremely reduced cellular and genomic organization. Here, microsporidia differ substantially from most other eukaryotes described so far. It is thus not surprising that it was initially tempting to equate organizational simplicity with a primordial primitive state, and thus to place microsporidia at the base of the eukaryotic tree. Microsporidia have a wide range of genome sizes. *Encephalitozoon intestinalis* possesses with only 2.3 Mbp one of the smallest eukaryotic genomes described so far (Vivarès and Méténier 2001). This is only about half the size of the genome of the bacterium *Escherichia coli* (Corradi et al. 2010). In contrast, the genome size of *Anncaliia algerae* is with 23 Mbp 10 times as large (Belkorchia et al. 2008), and the genome sequence of *E.aedis* spans as many as 51 Mbp (Desjardins et al. 2015). However, in general, microsporidia genomes are compact, and they are largely devoid of minisatellite repeats and transposable elements (Agnew et al. 2003). Yet, microsporidian genomes clearly show eukaryotic characteristics such as multiple linear chromosomes or telomeres. Moreover, their likely placement as sister to the fungi suggests, that their simplistic organization is the effect of a massive loss of cellular complexity. Together this has resulted in microsporidia becoming model organisms for studying reduction in eukaryotic genomes and metabolomes in the course of evolution (Williams and Keeling 2011; Wiredu Boakye et al. 2017).

The currently published microsporidian gene sets harbor only between 1,700 and 3,300 protein coding genes, again considerably fewer genes than many bacterial species (Heinz et al. 2012; Nakjang et al. 2013). Given this low number, it is likely that these genes approximate a minimal set that is essential for their life style as an obligate intracellular parasite (Agnew et al. 2003; Nakjang et al. 2013). Compared to their orthologs in other eukaryotes, microsporidian genes are mostly shorter (Katinka et al. 2001). They are flanked only by very short intergenic spaces, have few introns, and are poor in repetitive sequences (Keeling and Fast 2002; Corradi et al. 2010). Moreover, some of the genes are overlapping with each other (Corradi et al. 2010).

The diversity in genome sizes along with the varying, but consistently small, number of genes in microsporidia is thought to be the result of a complex evolutionary process including both reduction and expansion during the adaptation to their obligate intracellular parasitic lifestyle (Agnew et al. 2003; Williams 2009; Nakjang et al. 2013). Already the sheer extent to which the gene sets have shrunken on the microsporidian lineage suggest that these taxa heavily depend on their capability to tap their host's metabolism (Katinka et al. 2001; Luallen et al. 2016). Most prominently, microsporidia have lost their mitochondria alongside many biosynthesis pathways that are typically considered essential for life. Contemporary species seems to produce ATP solely via glycolysis instead of the more efficient Krebs cycle (Keeling and Fast 2002; Keeling and Corradi 2011; Heinz et al. 2012). To supply them, nonetheless with sufficient energy, microsporidia have established a dedicated transport system that can uptake ATP from their host species (Dolgikh 2000; Keeling and Corradi 2011; Heinz et al. 2012). Furthermore, it seems as if microsporidia are incapable of synthesizing purine and pyrimidine *de novo* (Heinz et al. 2014; Dean, Hirt, and Embley 2016). They seem to lack genes for several enzymes that are required to produce essential initial substrates for the purin and pyrimidine synthesis. In particular, the ribose-phosphate pyrophosphokinase that create phosphoribosyl pyrophosphate (PRPP), the IMP cyclohydrolase that synthesizes inosine monophosphate IMP, and the UMP synthetase that create UMP from PRPP are missing. It appears that microsporidia are using their nucleotide transport proteins NTTs, not only for the uptake of ATP, but rather to supplement the general nucleotide pool with resources taken from the host. Other key metabolic functions are shown to be missing in microsporidia such as the F_0F_1-ATPase complex, fatty acid synthesis or the formation of peroxisomes (Katinka et al. 2001; Cuomo et al. 2012).

1.5 The need for a deeper understanding of microsporidia

Microsporidia is a group of parasites that infect a wide range of species, many of which play important role in the agricultural economics as well as cause many medical troubles, especially infectious diseases in human. A deep understanding of the parasitic lifestyle and the physiology of microsporidia are required to better cope with this pathogen (Kaya and M. 2012; Bjørnson and Oi 2014). However, microsporidia are not only interesting because they are pathogens. In addition to the threat caused by microsporidiosis, microsporidia are a challenge to biologists who aim at unraveling their evolutionary origins and their evolutionary history. Furthermore, microsporidia with their compactness have been shown to be an exceptional model for studying the minimal eukaryotic genome as well as the obligate endoparasitic lifestyle in eukaryotic domain (Reinke and Troemel 2015). It is, thus, unfortunate that the particularities of microsporidian evolution and of their metabolism are still considerably poorly understood (Heinz et al. 2012; Nakjang et al. 2013). Research has been mainly hindered by two aspects: First, their lifestyle as an obligate intracellular parasite is a substantial obstacle to any experimental access. Any purified sample containing only microsporidia isolated from their host can contain only the microsporidian spores (Méténier and Vivarès 2001). Yet, the physiology of the sporal stage is thought to be substantially different from the developmental stages inside the host cell (Dolgikh, Sokolova, and Issi 1997). Furthermore, technique for genetic manipulation in microsporidia still faces many issues (Reinke and Troemel

2015). This has, so far, prevented the establishment of a microsporidian model system. At the same time, the tremendous evolutionary rates of microsporidian proteins (Slamovits et al. 2004), renders the inference of homology relationships to experimentally characterized proteins in other model organism hard. As a consequence, any *in silico* functional annotation transfer to microsporidian proteins suffers from a substantial lack of sensitivity (Jain, Haeseler, and Ebersberger 2018).

1.6 Outline of this thesis

The investigation of the microsporidia ancestor and novel approaches for a better functional annotating in term of accuracy and sensitivity could give more insights about the parasitic lifestyle as well as the metabolic system of the microsporidia, and can therefore provide a better knowledge for developing effective treatment methods against this emergent pathogen. Moreover, study the compactness of the microsporidian last common ancestor can help to understand the origin of the microsporidian reduction, which is either an ancestral state or only a secondary process that happened independently on the various contemporary species. By that we can approach a suitable method for designing the model for microsporidia.

In this thesis, we set out to shed further light on the evolutionary trajectory that molded the contemporary microsporidian pathogens from their last common ancestor. In a comparative genomics approach, we integrate the results of eleven microsporidian genome sequencing projects undertaken by the Microsporidian Genomes Consortium at the Broad Institute (https://www.broadinstitute.org/fungal-genome-initiative/microsporidia-genome-sequencing) and by the 1000 Fungal Genomes project of the Joint Genome Institute (https://genome.jgi.doe.gov/pages/fungi-1000-projects.jsf). In Chapter 2 of this thesis, we approach the problem of how to make the results of comparative gene set analyses provided by phylogenetic profiles amenable to an intuitive data analysis. To this end, we have developed a software, PhyloProfile, that facilitates an visual exploration of phylogenetic profiles, which can harbor, next to the presence-absence pattern of genes in individual genetic lineages additionally two information layers. In Chapter 3, we then propose a new solution of how to functionally annotate proteins, HamFAS. Here, we combine a targeted ortholog search that informs about the precise evolutionary relationships of the analyzed proteins, with an analysis of the feature architectures of already functionally annotated orthologous groups. In Chapter 4, we put the evolutionary analysis on microsporidia on a solid basis by first, pursuing a phylogenomics approach to establish a robust phylogeny of microsporidia and their placement in the eukaryotic tree of life. Then we perform a reconstruction of the gene set of the last common ancestor of the microsporidia and investigate its metabolic capacities in comparison to that of the contemporary species. Here we attempted to get insights into the evolutionary process of microsporidia for adapting to the obligate endoparasitic lifestyle as well as the origin of their genome and metabolism reduction.

Throughout this study, we analyzed the sequences in the amino acid level. Therefore, we treated the term *protein* and *gene* as synonym.

2 PhyloProfile: an interactive visualization tool for exploring complex phylogenetic profiles

2.1 Introduction

In evolutionary biology, the presence/absence pattern of a gene, the seed, across several species is defined as its phylogenetic profile (Pellegrini et al. 1999). Quantifying similarity between profiles gives an insight into co-evolving genes and thus can be used to transfer functions between genes (Jothi, Przytycka, and Aravind 2007; Date and Peregrín-Alvarez 2008). Moreover, phylogenetic profiles are commonly used for tracing gene clusters or biological pathways across species and time (Li, Calvo, et al. 2014; Dey et al. 2015; Wang et al. 2017). Although the evolutionary relationship is the basal information for phylogenetic profiling, it is not always informative enough to confirm the functional equivalence between two orthologs (Studer and Robinson-Rechavi 2009). For a more extensive profiling, the binary representation of genes is commonly integrated with additional information layers such as sequence similarities, similarities of domain architectures, or more general of feature architectures (Koestler, von Haeseler, and Ebersberger 2010) or semantic similarity of Gene Ontology-terms (Kensche et al. 2008).

Currently, there are few resources and tools available for such enriched phylogenetic profiles. Among these, DoMosaics (Moore et al. 2014) allows the annotation of individual groups of homologous sequences with features captured in profile HMM models, as they are provided e.g. by Pfam (Finn et al. 2016), The annotation results are then displayed graphically in the form of feature architectures. The user has the option to order the sequences and the corresponding architectures along a custom tree, facilitating the analysis of feature architecture change over evolutionary time scales. While DoMosaics is a valuable tool for displaying linear feature architecture for one group or homologous sequences, it has its considerable limitations. First and foremost, phylogenetic profiles of more than one protein family cannot be displayed. Moreover, only one type of feature can be displayed, and the feature architecture is strictly linear and does not allow overlapping features. The recently published Aquerium (Adebali and Zhulin 2017) is similar to DoMosaics, in its capability to display phylogenetic profiles alongside the feature architectures of the corresponding protein along a phylogenetic tree. Compared to DoMosaics, however the display options are substantially more comprehensive, and Aquerium can display simultaneously the phylogenetic profiles of up to ten proteins. This tool provides a scalable solution for the problem of visually inspecting and analyzing phylogenetic profiles. Unfortunately, it is almost completely devoid of any analysis functions that help in dynamically exploring, filtering, and interpreting phylogenetic profiles. The ETE3 tool kit (Huerta-Cepas, Serra, and Bork 2016) is a python framework for the display and analysis of phylogenetic trees. ETE3 provides options for displaying various annotations of the proteins at the leafs of the trees, among them also linear feature architectures. However, again this framework has been designed for displaying data rather than analyzing it. While few further tools may exist that can aid in the display and analysis of phylogenetic profiles, we are unaware of any software solution that facilitates the simultaneous display of multi-layered phylogenetic profiles containing hundreds or thousands of genes and taxa together with a comprehensive set of dynamically exploring and analyzing this data. Hence, we developed PhyloProfile, an interactive visualization tool for dynamically exploring such complex phylogenetic profiles.

2.2 Features and capabilities of PhyloProfile

PhyloProfile is a web browser based application to display information about the presence-absence pattern of one to many proteins across one to many species. Aside from visualizing the binary status of the orthologs, PhyloProfile can further represent multiple orthologous proteins for one search species, which are defined as co-orthologs. In these cases, a certain seed protein is represented by two or more sequences in a species that are more closely related to each other than they are to the seed. Representation of co-orthologs in the phylogenetic profile enables the exploration of whole genome duplication or the independent duplications of some certain genes. Next to the display of the phylogenetic profiles in a dot matrix, where a dot represents the presence of a protein in a particular species, PhyloProfile provides various options to dynamically filter the data depending to the additional information layers (variables) and the selected taxonomy rank. For example, minimizing the fraction of species required in a systematic group having a particular ortholog present can reduce the impact of spurious ortholog identification on evolutionary interpretations. Similarly, increasing the similarity cutoff for the protein feature architecture (Koestler, von Haeseler, and Ebersberger 2010) can help to filter genes that having divergent domain annotations. Once the filtering and data sub-selection is completed, the remaining data can be exported to serve as input for downstream analysis, such as phylogenetic and phylogenomic analyses.

PhyloProfile is intuitive, easy to handle and bridges the methodological gap between the gene set wide generation of phylogenetic profiles, e.g. via orthology prediction methods, and more focused downstream analyses concentrating on individual evolutionary or functional questions. The tool was written mainly in R (R Development Core Team 2011) with an intensive use of the Shiny library (https://CRAN.R-project.org/package=shiny).

2.2.1 Multiple input options

The data upload and the basic configuration of PhyloProfile is controlled by the Input & Settings page shown in Figure 2-1. Main input file for PhyloProfile is the phylogenetic distribution of orthologs or homologs for a set of seed proteins. For the sake of brevity, we will refer from here only to orthologs. The application extends similarly to homologs. The presence-absence patterns of orthologs can be complemented with up to two additional information layers, such as feature architecture similarities between related proteins, their sequence similarities, the evolutionary distances between the seed proteins and their orthologs, or any other measure between pairs of proteins. The main input file can be in tab-delimited text or multiple FASTA format. OrthoXML format (Schmitt et al. 2011) is also supported, however, its use is limited to the standalone version of PhyloProfile. In the main input file, the species has to be represented by their NCBI taxonomy IDs (Federhen 2012).

Next to the enriched phylogenetic profiles, the software can additionally display the feature architectures of the seed proteins and its orthologs/homologs. This information can be optionally uploaded into PhyloProfile in tab-delimited format.

If the user is interested in having the protein sequences ready for display as well, this information can be uploaded into PhyloProfile in standard FASTA format.

Alongside the data upload, a number of basic configurations can be set via the Input & Settings page. Further details are provided in the caption of Figure 2-1.

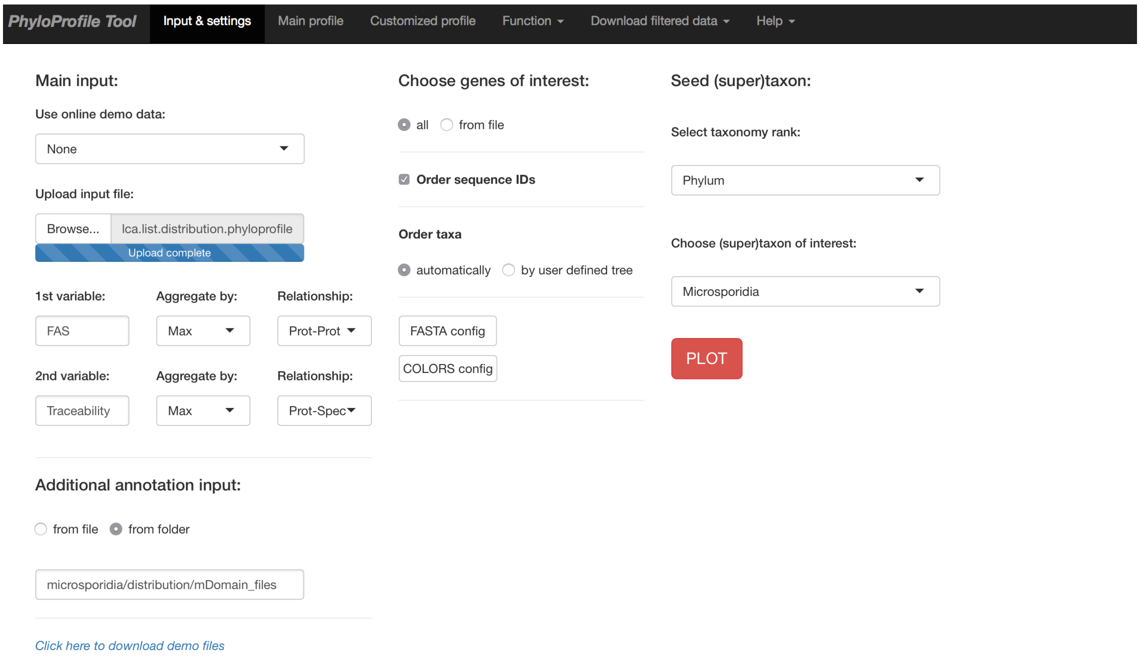

Figure 2-1: *Input & Settings* page of PhyloProfile. Users can upload here the main phylogenetic profiles, and optionally the domain annotation file and/or the sequence information in FASTA format. If the two additional information layers (variables) are not identified in the main input file, users can name them manually. Via this page, the user can additionally specify which subset of genes in the phylogenetic profile should be displayed. To this end, a text file with the genes of interest can be uploaded. Next, the user can choose to order the taxa automatically based on their systematic relationships as specified in the NCBI taxonomy. Alternatively, a custom tree in newick format can be uploaded that is then used for ordering the taxa. After adapting the default colors of the profile plots (if needed), users can select the taxonomic resolution, i.e. the taxonomic rank at which the phylogenetic profiles should be combined, ranging from strain to kingdom. Eventually, a taxon of interest can be selected that is displayed first in the graphic visualization of the phylogenetic profile.

Next to the use of custom phylogenetic profiles together with the accompanying information, PhyloProfile offers accompanying scripts for retrieving orthologous proteins for a set of seed proteins, together with their sequences and domain annotations, directly from the OMA Database using their REST-API (Altenhoff et al. 2015); likewise we have developed a set of scripts for parsing the outputs from OMA standalone (Train et al. 2017), hmmscan (hmmer.org) and pfamscan (Finn et al. 2014) to generate the compatible inputs for PhyloProfile.

The last required input information is the systematic taxonomy rank for the analysis and the corresponding reference taxon, which can be selected from the Input & Setting page of the tool (Figure 2-1).

2.2.2 The use of NCBI taxonomy information in PhyloProfile

In order to achieve a more meaningful interpretation of evolutionary events from the phylogenetic profile, the taxa given in the profile have to be arranged such that the time to the last common ancestor shared with the reference taxon increases from left to right, assuming that the x-axis of the profile plot represents taxon list. The order of taxa can be inferred from the a taxonomy tree that reflects the overall species phylogeny (Hinchliff et al. 2015). There existed a function `class2tree` in the R library taxize (Chamberlain and Szocs 2013), which could reconstruct a taxonomy tree from a list of taxa by the use of the NCBI taxonomy information for the main defined ranks, such as strain, species, genus, up to the superkingdom. However, by excluding the undefined ranks, which are named as "norank" by NCBI, the tree resulted from this function became multifurcating, especially for clades of close related taxa. Consequently, the order of taxa in the multifurcate clades is incorrect. We therefore developed a new approach for better resolving the reconstructed taxonomy trees. This approach has been implemented in the new version of the taxize library (Chamberlain et al. 2018).

First, we collect the full NCBI taxonomy information for a list of input taxa including both defined and undefined ranks. Not in all cases, all taxonomic ranks that are represented in the NCBI taxonomy are specified for each species represented in the NCBI taxonomy. It is for this reason that the vector of taxonomic information can differ in length between the individual species. To make the information consistent across all species, we implemented an R function to align the individual taxonomy vectors. This results in a taxonomy matrix, in which the rows represent the taxonomy IDs and the columns are all available systematic ranks that can be found in the given taxon list. If a certain species has no information displayed for a rank, the value in the corresponding cell is obtained from the previous defined rank. On the basis of this taxonomy matrix we then generate a taxonomy tree using the modified *class2tree* function of the taxize library. Thereafter, we root the tree based on the user-selected reference taxon and return a list of sorted taxa from the rooted tree. This feature facilitates the analysis from individual species to classes, phyla or entire kingdoms. Note, our software allows to integrate taxa, which are not yet represented in the NCBI taxonomy database, into this process.

2.2.3 Interactive visualization

The Shiny library (https://CRAN.R-project.org/package=shiny) enables the interactive visualization in PhyloProfile. Interactive visualization has show its robust ability in analyzing informative data (Zudilova-Seinstra, Adriaansen, and van Liere 2009).

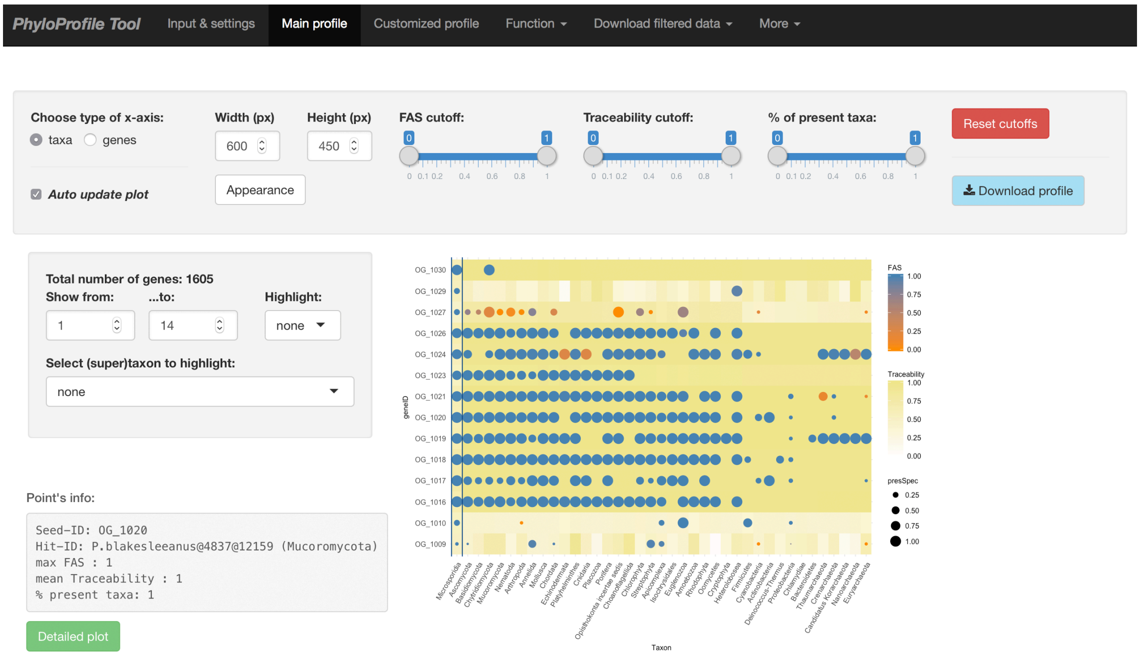

Figure 2-2: The *Main profile* page of PhyloProfile. The phylogenetic profile is represented by a dot matrix. Cell color and dot color denote values of two additional information layers. The Dot size is proportional to the fraction of species summarized at the selected taxonomic level (cf. Figure 2-1) that harbor at least one ortholog to the seed. The assignment of taxa and genes to the x and y-axis, respectively, can be switched if desired. Co-orthologs to a seed in a given taxon will be represented as a

small green dot inside the main dot. This feature is available is only when the taxonomic resolution is set to *species*. The detailed information for a seed – ortholog pairing (lower left part of the plot) can be accessed upon a click on the corresponding dot in the dot matrix.

Once all information is uploaded into PhyloProfile, the phylogenetic profile can be accessed via the Main profile page (Figure 2-2). This page harbors the main profile as a dot matrix. Depend on the selected taxonomy rank, the dot size can be vary. If the selected rank is the most specific one based on the input data, namely species or strain in the most cases, the co-orthologs can also be represented by another small dot inside the main dot. The size of the internal dots indicates the number of co-orthologs present in a species (Figure 2-9).

Additionally, several options exist to customize the plot. Options range from the simple adaptation of the plot layout, via the specification of the number of genes that should be displayed in the dot matrix, up to filtering options that consider the additional information layers provided alongside the phylogenetic profile. By specifying individual cutoff values the user can blend out all entries in the dot matrix that do not meet the filtering criteria. For example, it is possible to remove all orthologs whose feature architecture similarity to the seed protein is below the specified limit. Upon a click on a dot in the matrix, the detailed information about the underlying seed – ortholog pair can be accessed (Figure 2-3).

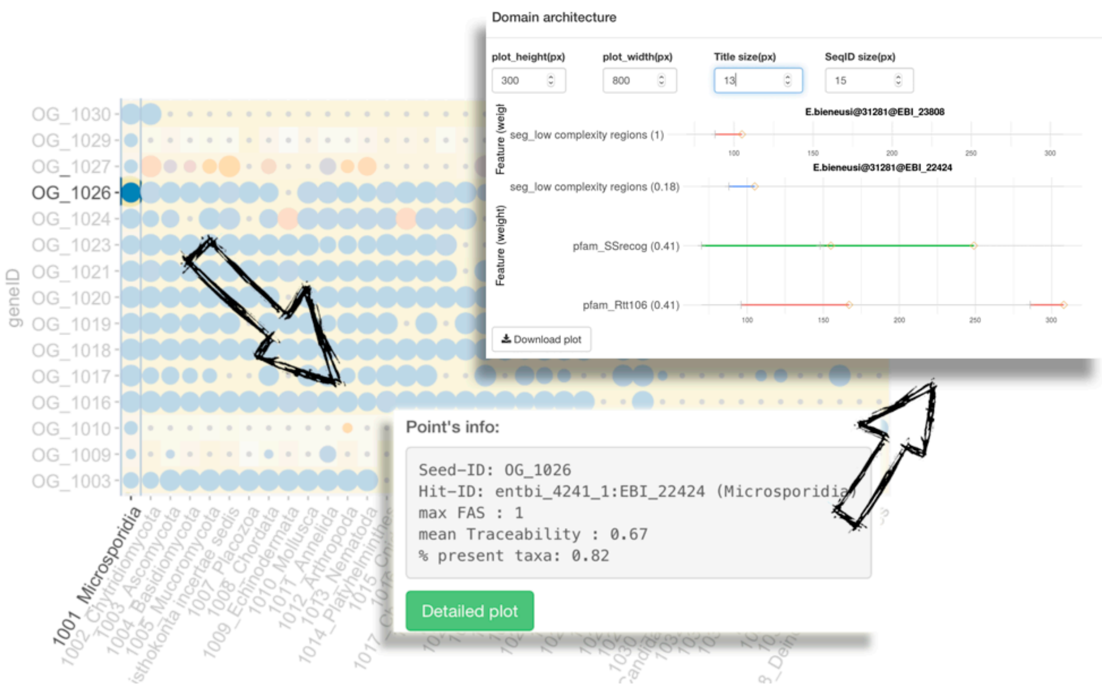

Figure 2-3: The interactive visualization enables a rapid adaptation of the focus to the desired level of resolution. The information stored in PhyloProfile ranges from the overview image of the phylogenetic profiles of hundreds to thousands of proteins and species down to the pair-wise analysis of feature architectures.

All plots generated in PhyloProfile are interactable in order to represent further data or to link between different functions as described in section 2.2.5 below.

2.2.4 Subselecting taxa and genes via the Customized profile page

The main profile page shown in Figure 2-2 is designed to display either the full phylogenetic profiles, or slices therefore by defining a consecutive set of genes for display. A selection of individual genes and species is not possible. This feature is accomplished in the customized profile page of PhyloProfile. Here, a detailed analysis of a subset of genes and taxa, without the need of modifying the input data, is possible. The genes and taxa used in customized profile (Figure 2-4) can be manually selected from a pre-defined list or can be imported directly from the various analysis functions provided by PhyloProfile (see section 2.2.5 below).

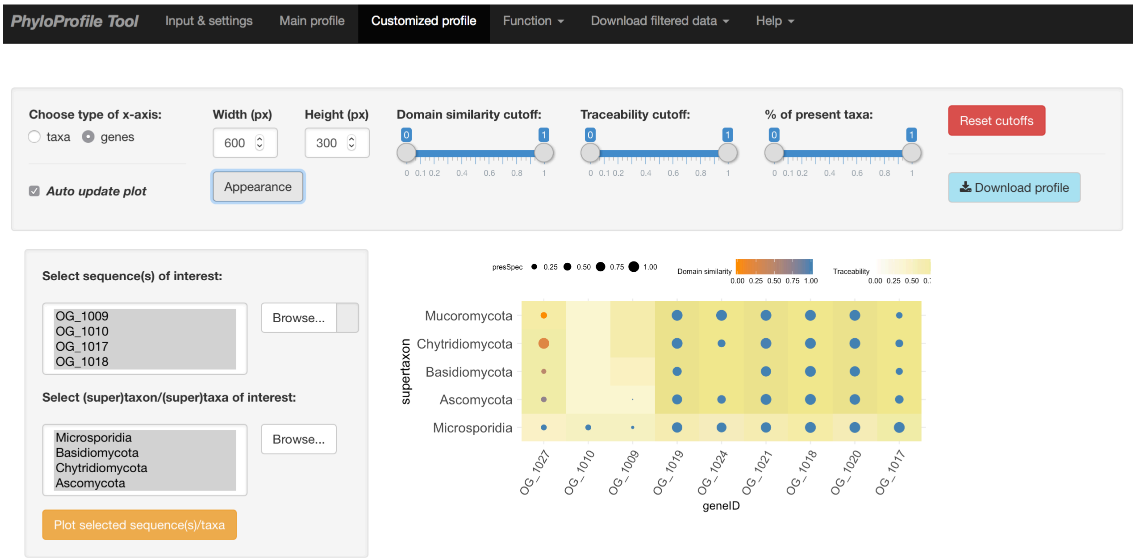

Figure 2-4: Customized profiles of 9 selected proteins in microsporidia and 4 chosen fungal phyla.

2.2.5 Analyzing phylogenetic profiles

In addition to the interactive visualization, PhyloProfile further provides several functions for dynamic analyzing the phylogenetic profiles.

Profile clustering

The similarity of phylogenetic profiles can be an evidence for the functional relation between proteins (Pellegrini et al. 1999; Jothi, Przytycka, and Aravind 2007; Date and Peregrín-Alvarez 2008). We, therefore, create a function to cluster genes according to the distance of their phylogenetic profiles (Figure 2-5).

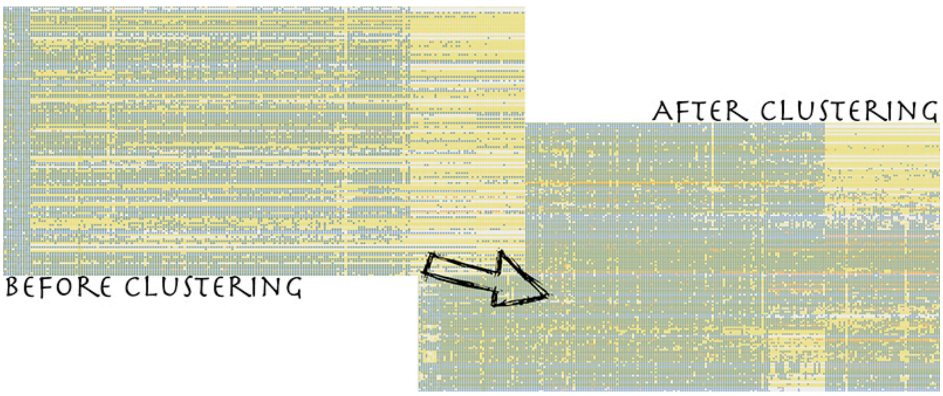

Figure 2-5: Phylogenetic profile dot matrix before (left) and after (right) clustering the proteins according to the distance of their phylogenetic profiles. The clustered profile clearly reveals the existence of three main groups of genes that differ in their phylogenetic distribution.

First, a distance matrix will be generated from the pairwise distances between all presence/absence profiles with the `dist` function. Then, we cluster the similar profiles based on a chosen clustering algorithm from the `hclust` function in R. An overview of the currently implemented distance matrix measures and clustering algorithms is shown in Table 2-1.

Table 2-1: Implemented distance matrix measures and clustering algorithms

Distance matrix measures from `dist` function	Clustering algorithms from `hclust` function
asymmetric binary	single linkage
canberra	complete linkage
euclidean	average (UPGMA)
manhattan	mcquitty (WPGMA)
maximum	median (WPGMC)
	centroid (UPGMC)

Gene age estimation

Phylogenetic profiles provide information about the evolutionary age of the genes under study, as has been shown in the study of the ribosome biogenesis pathway of Ebersberger et al. (2014). PhyloProfile lets the user assess the evolutionary age of each gene in the phylogenetic profile using an LCA (last common ancestor) algorithm (Capra et al. 2013). Namely, the last common ancestor of the two most distantly related species in the ortholog group serves as the minimal gene age of that group. To accomplish the LCA inference in PhyloProfile, we implemented the function to find the most remote taxon to the reference species in each ortholog group. Using the reconstructed taxonomy tree as described in 2.2.2 above, we identified the LCA for those two taxa and assigned the evolutionary age for the corresponding ortholog group by that LCA. Figure 2-6 shows an example result of the gene age estimation routine.

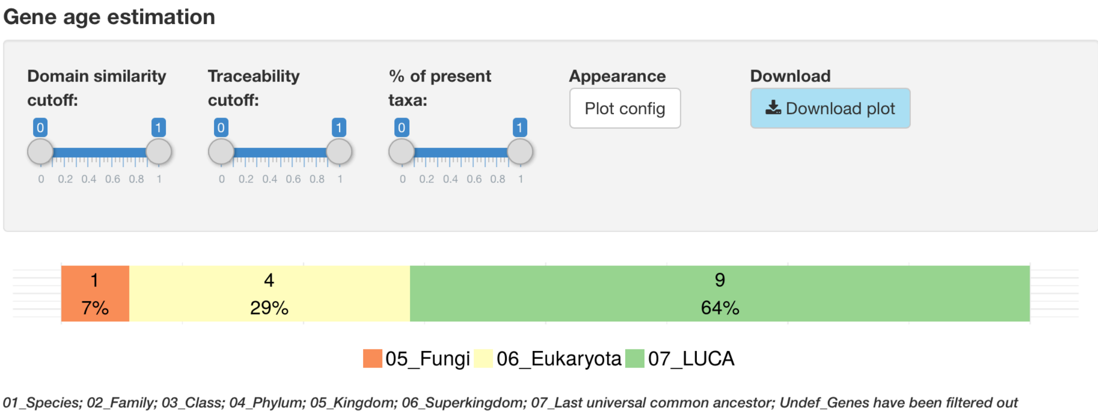

Figure 2-6: Gene age estimation based on LCA algorithm. The different colors in the age profile denote the individual gene ages. The numbers of genes subsumed in each age layer are given within the colored areas together with the percentage of the total gene set. The age layer is interactive. Upon a click on either layer, the corresponding list of genes is displayed (not shown). The information can then be either displayed in the Custom profile page, or downloaded as text file.

Core gene identification

Core genes are genes that are shared among all taxa in a user specified group. One routine use of core gene compilations is the reconstruction of phylogenetic trees integrating the phylogenetic trees of as many genes as possible while, at the same time, minimizing missing data due to absent genes (Daubin, Gouy, and Perrière 2002). We have implemented a routine into PhyloProfile, which lets the user define a set of taxa in which a gene has to be present in order to be assigned a core gene status. The resulting core gene set can be displayed in the Customized profile page and optionally downloaded together with the corresponding sequence data.

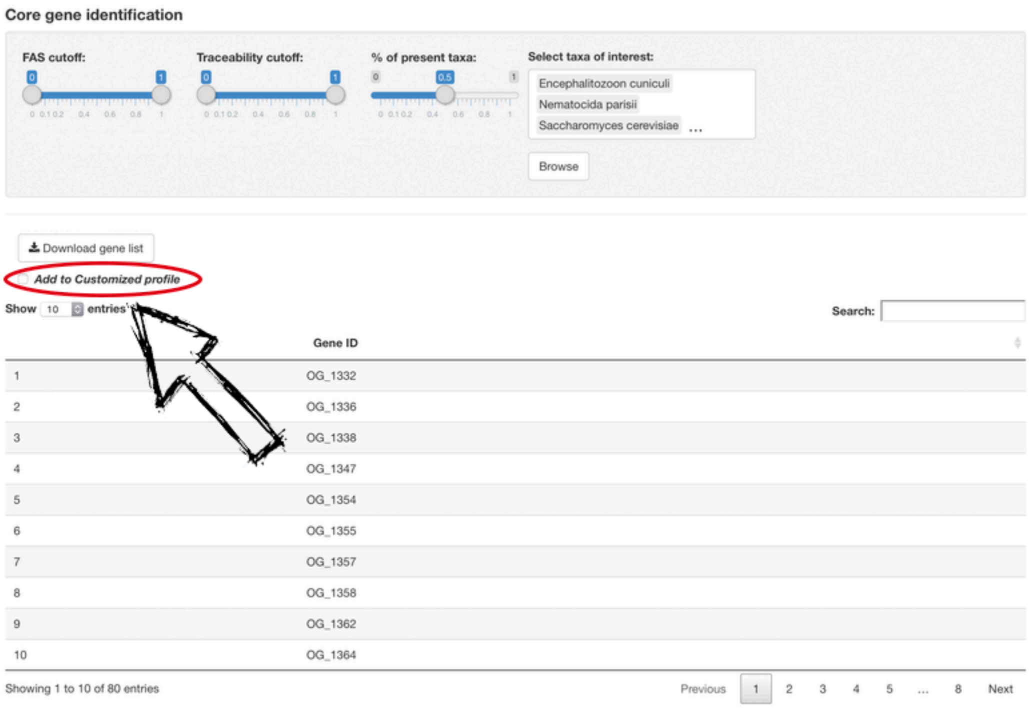

Figure 2-7. List of genes resulting from the *Core gene identification* function can be directly input to the customized profile for further investigating.

Distribution analysis

One of the main innovations of PhyloProfile is the enrichment of phylogenetic profiles with accessory information that provide information, e.g. about the similarity of the orthologs to the seed. To explore the distribution of these additional measures across the phylogenetic profile, we have implemented the Distribution analysis function (Figure 2-8). Here, the distribution of the values of up to two integrated information layers – if provided - and the percentage of species summarized at the selected taxonomic resolution can be visualized. Next to informing about the general distribution of these values across the data, the plots can be used for outlier detection, e.g. proteins with a lower extent of similarity to the seed when compared to the other orthologs. This helps to decide on filtering threshold for a downstream data analysis.

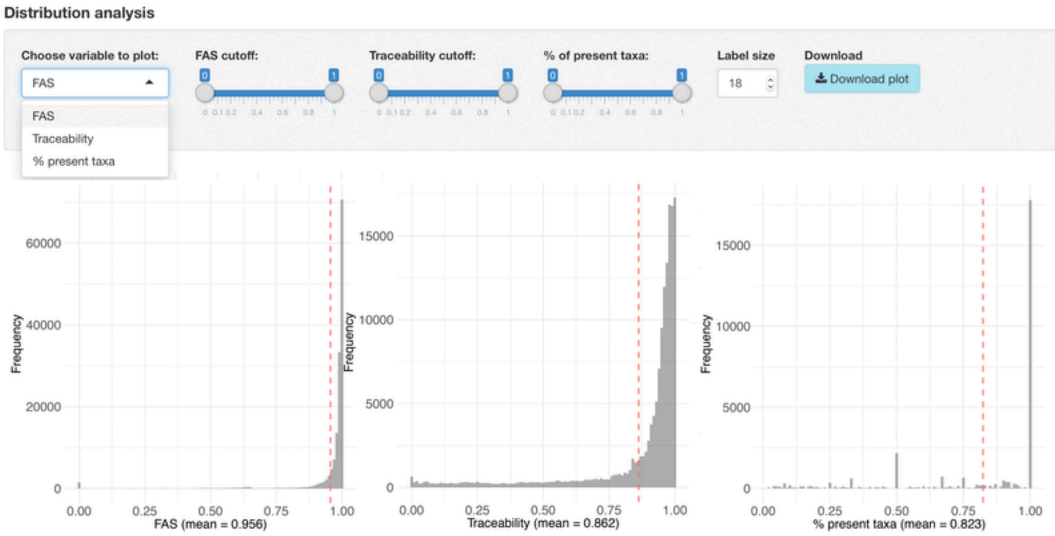

Figure 2-8: Distribution analysis of two integrated data and the fraction of species in the systematic group. Those distributions can be dynamically changed depending on the defined thresholds of those variables.

2.2.6 Interoperable output

All plots generated in PhyloProfile can be exported as PDF files. Filtered profiles as well as their sequences can be downloaded as a list and multi-FASTA file for further downstream study, such as phylogenetic tree reconstruction or metabolic pathway analysis.

2.3 Result

2.3.1 Availability of PhyloProfile

PhyloProfile is distributed with an exhaustive documentation (https://github.com/BIONF/PhyloProfile/wiki) and several testing data sets (https://github.com/BIONF/phyloprofile-data/tree/master/expTestData). Figure 2-9 represents the phylogenetic profile of the AMPK-TOR pathway (Roustan et al. 2016) testing set.

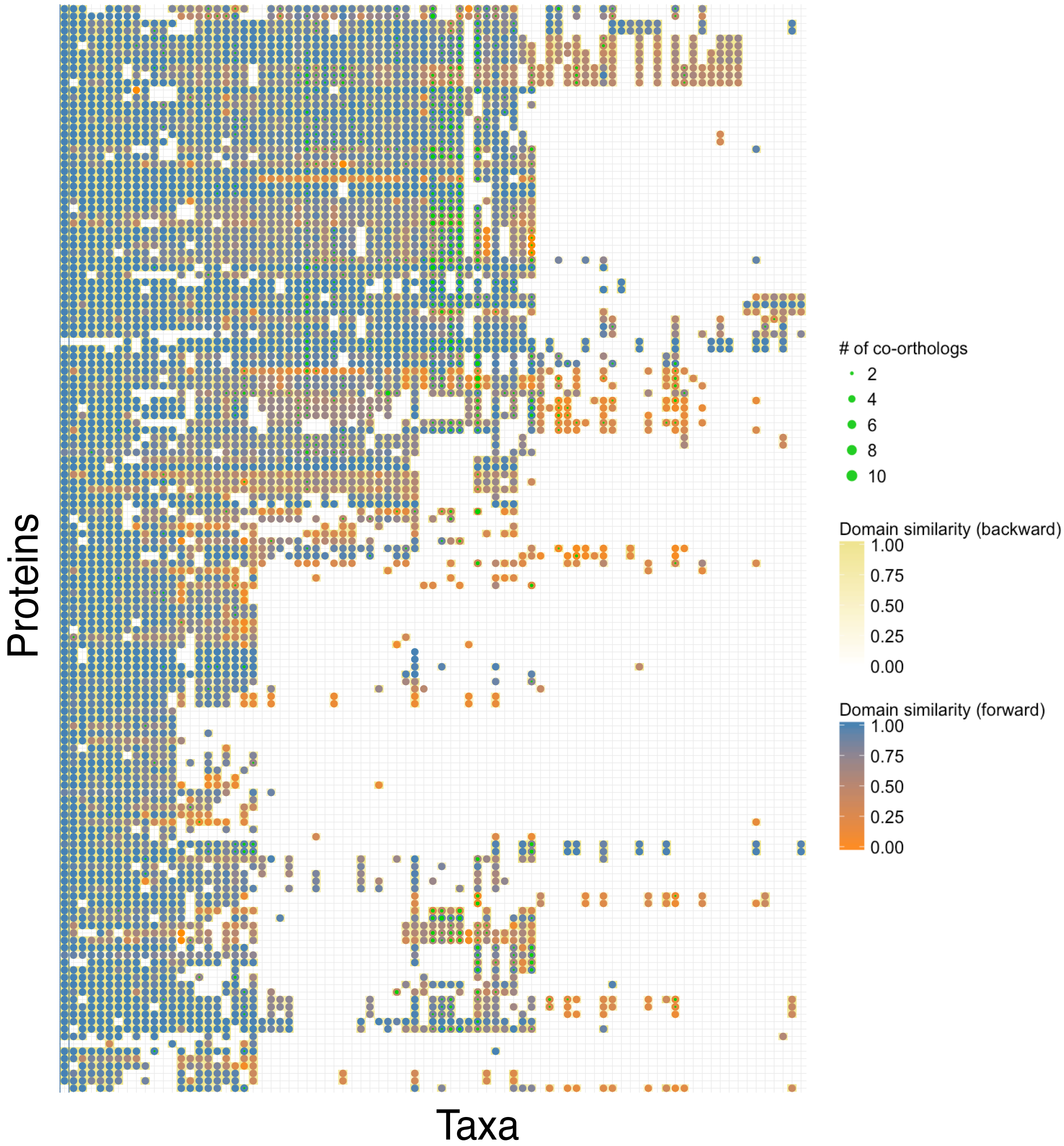

Figure 2-9: Phylogenetic profile of AMPK-TOR pathway.

The standalone version as well as the open source code of PhyloProfile can be found at https://github.com/BIONF/PhyloProfile/releases. Besides, we also offer an online version at http://applbio.biologie.uni-frankfurt.de/phyloprofile/, which can run directly on the web browser without any installation.

2.3.2 Performance test

We checked the performance of PhyloProfile by assessing the time required for both importing and plotting the full data (Figure 2-10), and RAM usage (Figure 2-11) with different data size. As test data served the phylogenetic profiles of

1,605 microsporidian proteins across 489 species. The full data matrix comprises 784,845 cells. It takes about 70 seconds to load the data and about 180 seconds to plot the entire matrix. We then reduced the data matrix stepwise by either considering fewer genes (Figure 2-10 a) or fewer taxa (Figure 2-10 b), and measured the time to upload and plot the data.

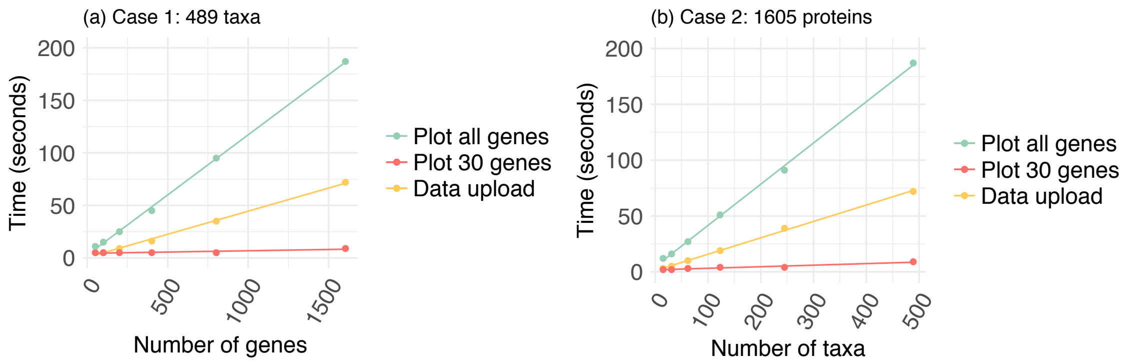

Figure 2-10: The running time of PhyloProfile for uploading (yellow) and plotting phylogenetic profiles of all (green) or the first 30 genes (red) scales linearly with data size. (a) Running time as a function of number of genes analyzed. (b) Running time as a function of number of taxa analyzed.

Plotting of the first 30 genes (default setting) is independent of the data size. The phylogenetic profile of a moderate sized data set comprising 200 genes and 200 species (40,000 cells) takes about 10 seconds to display, both on the standalone version and on the online version. The results indicate that PhyloProfile facilitates a reasonably quick interactive exploration of the data for data comprising up to a few hundreds of genes and taxa. We trust that this will be sufficient for the vast majority of applications, as we expect that a typical user will be interested in exploring phylogenetic profiles of gene sets representing, e.g. one or few KEGG pathways (Kanehisa et al. 2016). However, the analysis of substantially larger data is also possible, and the option to extract subsets of interest via the customized profile option allows streamlining and speeding up the analysis.

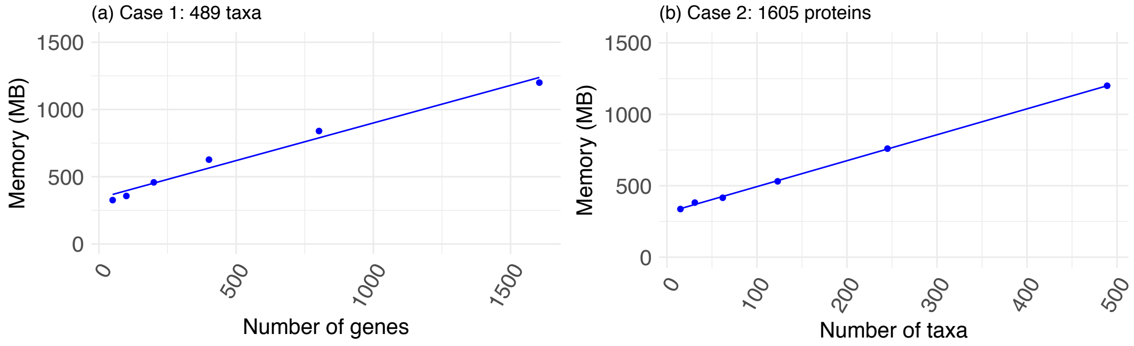

Figure 2-11: RAM usage during data display increases linearly as the data matrix grows. (a) RAM usage as a function of number of genes analyzed, and (b) as a function of the number of taxa analyzed.

An online version of PhyloProfile is available via the webserver of the Applied Bioinformatics group at http://applbio.biologie.uni-frankfurt.de/phyloprofile/ using the shiny server that is provided as a service to the community by RStudio Inc. The performance of the online version is comparable to the standalone version with respect to speed of data upload and plotting of the profiles. The only difference is, that the online version currently does not support orthoXML format as input.

2.4 Discussion

With PhyloProfile, we have now a novel software at hand, that allows to adjust the focus of comparative gene set analyses from overview analyses – where the phylogenetic profiles of hundreds to thousands of proteins across the same number of species can be analyzed – to high resolution analyses, where the feature architecture of individual protein pairs can be inspected. On top of displaying the presence/absence patterns of genes across the taxa under study, PhyloProfile is, to our knowledge, unique in its ability to display up to two additional information layers. It is thus possible to integrate the results from complementary analysis, e.g. the pair-wise comparison of protein sequences or feature architectures, into the phylogenetic profiles, such that all information is available at one glance. We trust that this facilitates a more meaningful interpretation of phylogenetic profiles in the context of the evolution of function. The dynamic filtering of the supplementary data such as the domain architecture similarity or the evolutionary distance between the seed proteins and their orthologs can help to reduce the impact of poorly orthology assignment based on sequence similarity. Furthermore, PhyloProfile was designed with interactive visualization ability. It facilitates the visualization and exploration of phylogenetic profiles together with the protein feature architectures in an interactive and effective way.

However, there is still room for improvements. First and foremost, the running time becomes an issue in the light that novel genomes – and the encoded gene sets – emerge nowadays almost on a daily basis. Although PhyloProfile can handle large phylogenetic profiles, its performance is not yet sufficient for interactively displaying and analyzing data comprising entire gene sets of a typical eukaryote (10,000 genes and beyond) and thousands of species. While an a priori filtering of genes and the selection of representative taxa prior to the upload into the tool is a viable approach to cope with this problem, ultimately, a more efficient implementation speeding up both display and analysis will be necessary. Here, the re-implementation of PhyloProfile in a programming language other than the considerably slow R will be necessary. Besides, we are working out for implementing some further practical features, such as identify the convergence point of two orthologous proteins, which is the time when one gene got duplicated into two copies. For a thorough phylogenetic analysis, we are planning to create an automatic pipeline from searching orthologs, phylogenetic profile exploration to phylogenetic tree reconstruction and pathway analysis.

3 HamFAS: a novel functional annotation approach based on feature-aware orthology inference

3.1 Introduction

3.1.1 Functional annotation transfer

Over the past twenty years, the costs to determine the sequence of a eukaryotic genome have dramatically decreased by several orders of magnitude. The sequencing of the first eukaryotic genomes, such as those of *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Drosophila melanogaster* (Adams et al. 2000) or human (Lander et al. 2001; Venter et al. 2001) were multinational ventures whose costs were in the range of several hundred million US dollars (US-$). Meanwhile, a genome sequence of a human-sized eukaryotic genome is available for little more than one thousand dollars (Wetterstrand). In particular, the price for a mega base of raw sequence data has fallen below 0.1 US-$. At the same time, standard bioinformatics workflows have been established to ease the detection and annotation of the gene sets encoded in these genomes, e.g. MAKER (Cantarel et al. 2008) or AUGUSTUS (Stanke and Morgenstern 2005). However, the sheer identification of the individual genes does provide per se no information about the encoded function. Thus, to further annotate these genes, and to get insights into the spectrum of molecular functions represented in a species' genome further efforts are required. In essence, function assignment is one of the crucial steps in every sequencing projects to characterize the predicted genes or proteins in the new genomes (Gabaldón and Huynen 2004). In the pre-genomic era, the only way to assess a gene's function involved the experimental characterization of the gene product (Alberts et al. 2002b). And even nowadays, efforts are underway to inactivate every gene in the genome of individual model organisms to shed light on the likely function of affected gene (e.g. (Grimm 2006) or (Hall, Limaye, and Kulkarni 2009)).

The alternative way to assign function to an unknown protein is the *in*-silico transfer of functional annotation from one protein to another. *In-silco* transfers gained only then momentum, once comprehensive information about proteins and their functions were available in the public databases. Moreover, standardized approaches had to be developed to make this information accessible to computational analyses (e.g. GO (Ashburner et al. 2000) or KEGG (Kanehisa et al. 2016)). In general, we can distinguish between two main kinds of *in-silico* methods for functional annotation transfer, (i) structure-based approaches, and (ii) sequence-based approaches.

Three-dimensional structures provide detailed insights into the functions of proteins (Adams et al. 2007). As protein structures evolve exponentially slower than the underlying amino acid sequences, they should be particularly helpful in assessing the function of an unknown sequence (Chothia and Lesk 1986; Williams and Lovell 2009). Similar structures can therefore also be found between dissimilar sequences (Rost 1997). Additionally, protein's functions are identified mainly by the corresponding conformation (Laskowski 2009). Their structures are therefore more informative then the sequences regarding to protein's functionality (Friedberg 2006). However, structure-based annotation transfer methods have their limitation. Firstly, there is a huge gap between the number of protein structures and published sequences (Lee, Wu, and Zhang 2009). As of today only 145,000 structures have been deposited in the PDB (https://www.rcsb.org/stats), contrasted by 4 times as many curated sequences deposited in UniProtKB/Swiss-Prot (https://web.expasy.org/docs/relnotes/relstat.html), and almost a 1000 times as many un-curated sequences currently hosted by UniProtKB/TrEMBL (http://www.uniprot.org/statistics/TrEMBL). Secondly, a large fraction of available structures have no functional annotation (Nadzirin and Firdaus-

Raih 2012). Even more importantly, the protein structure prediction process is time consuming and complex (Baker 2001). Probably, one of the most challenging tasks is that a comprehensive modeling framework is missing, which helps to differentiate a significant deviation from a protein structure from a variation by chance that has no effect on the encoded function. Thus, structure-based annotation methods face challenges for automatic annotation of a large number of new sequenced proteins, especially in this era of whole genome sequencing (Watson and Thornton 2009). It is for all these reasons that a functional annotation transfer based on sequence similarity is still primarily used as the first step for assessing the function of a newly characterized protein (Sael, Chitale, and Kihara 2012).

The most common tools used for a functional annotation transfer exploit sequence similarity (Loewenstein et al. 2009). The fundamental rationale for of this approach is based on an evolutionary concept. Proteins displaying a sequence similarity that is higher than it is expected by chance are most likely homologous, i.e. they share part of their evolutionary ancestry and date back to a single common ancestral sequence. It is, thus reasonable to assume that they have similar functions (Gabaldón and Koonin 2013). One of the easiest approaches to accomplish a similarity based annotation transfer is based on database search heuristics, such as BLAST (Altschul et al. 1997). Here, a protein sequence with unknown function is used as query to search for functionally annotated homologous sequences in different sequence databases such as the non-redundant protein of the National Center for Biotechnology Information (NCBI Resource Coordinators 2017), or the UniProt Knowledgebase (Bateman et al. 2017). The functional annotation from the top hit sequence is used to tentatively annotate the query (Gabaldón and Huynen 2004; Friedberg 2006). While BLAST based approaches have the advantage of being quick and conceptionally simple, they carry the drawback of comparing only pairs of sequences. Moreover, the pure extent of sequence similarity is, in many cases, only a considerably poorly proxy for a functional similarity. This is, because the typically used scoring matrices, such as BLOSUM62 (Eddy 2004) in the case of BLAST, evaluate the similarity between proteins in a position-independent way. More precisely, a sequence difference at a functionally relevant site, e.g. a functional tyrosine residue in a tyrosine kinase, reduces the similarity score to the same extent than the same sequence difference at a position that is not relevant for protein function. It was not at last for this reason that dedicated databases have been established, such as the Protein Family Database Pfam (Finn et al. 2016) or SMART (Letunic and Bork 2018), where proteins – or parts thereof – that are evolutionarily conserved, and share the same function, are grouped into protein families. The sequences in each family are, in the case of Pfam, manually curated, aligned, and the resulting profile is then used to train a profile hidden Markov model (Eddy 1998). These models aim at capturing position specific sequence characteristics that are relevant for the function of the proteins in the training data. Search algorithms, such as hmmscan or hmmsearch from the HMMER package (Finn et al. 2015) can then be used to identify significant hits to these modeled domains in a query sequence. If such a hit is detected, the query can be annotated with the function represented by the pHMM. These profile-based approaches have shown a dramatically higher sensitivity than sequence-based search, especially for related sequences whose identity goes below 30% (Park et al. 1998).

3.1.2 Standardized description of protein function

Protein functions are mostly described using free text (Lee, Redfern, and Orengo 2007). Because the same function can be described with different but synonymous terms, the comparison of non-standardized functional annotations across proteins is a non-trivial problem. It can easily confuse humans, and it is even more challenging for computers, where algorithms have to decide whether the functional description of two proteins is similar or not (Friedberg 2006). To make the field of comparative protein function analyses accessible for automated *in-silico* analyses it was necessary to develop and implement a systematic language for describing the protein's functions, which is human friendly and computer-readable (Wilson, Kreychman, and Gerstein 2000; Lan, Jansen, and Gerstein 2002). The first famous effort was introduced by the biochemists who developed the Enzyme Commission (EC) classification (Webb 1990). A combination of four numerical blocks describes the enzymatic function of a protein with increasing precision. This classification system was designed explicitly for enzymatic functions. It lacks the conceptual flexibility that would have allowed its adoption for describing the plethora of functions of proteins other than enzymes. Here, the system of using controlled vocabularies organized in a hierarchical structure, so called ontologies, turned out as viable alternatives. In 2000, Ashburner et al. proposed the Gene Ontology (GO) for annotating gene products. GO classifies protein functions into three categories, biological process, molecular function, or cellular component. Their hierarchical structure is represented as a directed acrylic graph, which enables comparing annotations at different levels of precision.

KEGG orthology identifiers are also widely used to describe gene functions (Yin et al. 2016), yet with a higher granularity compared to the GO terms. KEGG, the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto 2000; Kanehisa et al. 2016), is a resource for sequence annotation and biological pathway analysis. It provides comprehensive information with 16 databases comprising of four groups: System information (with three databases PATHWAY, BRITE, MODULE), Genomic information (ORTHOLOGY, GENES, GENOME), Chemical information or KEGG ligand (COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS, ENZYME), and Health information (DISEASE, DRUG, DGROUP, ENVIRON). The hub to link all those databases is the KEGG Orthology (KO). Each KO entry, defined by a K number, comprises similar sequences, which are pathway dependent (Kanehisa et al. 2014). Originally, the KO database was developed together with KEGG pathway maps, BRITE functional hierarchies and KEGG modules based on experimental knowledge. Meanwhile, the assignment of K numbers to KEGG GENES using auto KOALA (KEGG Orthology And Links Annotation) algorithm is a highly automatized process. Manual KOALA checking is, however, still required if there are discrepancies between current annotations and new assignments (Kanehisa et al. 2016).

3.1.3 Functional annotation transfer between homologs

It is common practice to use homology relationships between proteins as a main argument for a functional annotation transfer (Loewenstein et al. 2009). Homologous genes, i.e. genes that share a common ancestry, can be divided into orthologs and paralogs. Orthologs are genes whose genetic lineages separated by a speciation event, while paralogs are the results of from a gene duplication event (Fitch 1970). It has been supposes, that genes after being duplicated can evolve freely and develop new functions (Ohno 1970). Therefore, orthologous sequences are thought to be more similar in function than paralogous sequences (Koonin 2005; Altenhoff et al. 2012; Chen and Zhang 2012; Thomas et al. 2012), that orthologs are indeed significantly more similar than paralogs with respect to their function. However, these and other studies convincingly have shown that also orthologs can differ in their function. For example, it was shown that sequences with high rate of similarity could have different functions (Rost 2002; Tian and Skolnick 2003). In contrast, the same function can also be retained in dissimilar sequences (Whisstock and Lesk 2003). Thus, orthology relationship needs to be combined with other evidences to facilitate a meaningful prediction of a protein's function. This is the more the case when the aim is to transfer functional annotations between distantly related species (Reid, Yeats, and Orengo 2007). Those supported evidences could be gene modules (Kachroo et al. 2015), co-expression network information (Bargsten et al. 2014), or the combination of both structure and protein-protein interaction data (Zhang, Freddolino, and Zhang 2017). Alternatively, the feature architecture similarity of the corresponding proteins can be assessed (Koestler, von Haeseler, and Ebersberger 2010). A feature architecture is the arrangement of different types of protein domains such as Pfam, SMART domains, transmembrane domains or low complexity regions. The comparison of feature architectures between two proteins gives a FAS score between 0 and maximally 1.

3.1.4 KAAS and BlastKOALA

KEGG provides two online annotation servers for assigning K numbers to query genes, and both rely on the identification of homologs. The first introduced tool was KAAS - KEGG Automatic Annotation Server (Moriya et al. 2007). This approach is based on bidirectional BLAST searches against a set of user-defined reference species. A ranked score calculated from the BLAST hits serves as the decision criterion if a KO identifier is assigned to the query gene, or not.

Recently, a new annotation server of KEGG, the BlastKOALA has been published (Kanehisa, Sato, and Morishima 2016). This modified version of the KOALA algorithm is used to assign K numbers to query sequences. It is based on the GFIT (Gene Function Identification Tool)-like table converted from the results of the BLAST search of the query against a collection of reference proteins. KOALA computes the weighted sum of scores from the unidirectional BLAST searches. The factors determining the weight are the length of the alignment, the ratio of query and target sequence lengths, the degree of matches of taxonomic categories (if known), and the degree of matching Pfam domains. In addition, a number of further rules are checked to identify the best fitting K number for the query sequence. Different from KAAS, BlastKOALA uses a non-redundant dataset created from GENES database as reference for the BLAST search. Kanehisa, Sato, and Morishima (2016) showed exemplarily for the genome of the bacterium *Kangiella geojedonensis*, that with the non-redundant dataset and the new KOALA algorithm, BlastKOALA outperformed KAAS.

3.1.5 The need for a novel sequence-based annotation transfer approach

The existing KO annotation tools KAAS and BlastKOALA, though they are widely used, have their limitations. For example, KAAS infers annotations based only on reciprocal BLAST searches. While this approach serves to identify orthologs among pairs of species, it does not make use of further evidences, which can lend additional support to a functional annotation transfer. In turn, BlastKOALA performs only a unidirectional blast, which leaves the precise nature of the evolutionary relationships between query and hit in the dark. However, in contrast to KAAS, BlastKOALA considers further supporting information to improve the accuracy of the annotation transfer. Unfortunately, the underlying weighting scheme as well as other rules for identifying the best hit remain largely unclear (Kanehisa, Sato, and Morishima 2016). Furthermore, BlastKOALA is limited at maximum 5,000 to 10,000 sequences per job, which makes an automatic annotation for a larger set of sequences hard.

Optimally, a software for functional annotation transfer would make full use of both the precise determination of evolutionary relationships among sequences and the comparison of their feature architectures. Here, we describe HamFAS, a novel annotation approach based on feature-aware orthology inference. In a nutshell, HamFAS performs a targeted ortholog search for the query protein. Subsequently the orthology assignments are filtered based on the feature architecture similarity (FAS) (Koestler, von

Haeseler, and Ebersberger 2010) between orthologous proteins and their seeds. HamFAS automatically infers a KO-specific FAS score cutoff by initially determining the pair-wise FAS scores within a KO group.

3.2 The HamFAS approach

3.2.1 Algorithm

HamFAS is a hybrid approach for an annotation transfer of KO ids, which integrates a targeted ortholog search and the assessment of feature architecture similarities between ortholog pairs. Figure 3-1 informs about the general workflow of HamFAS.

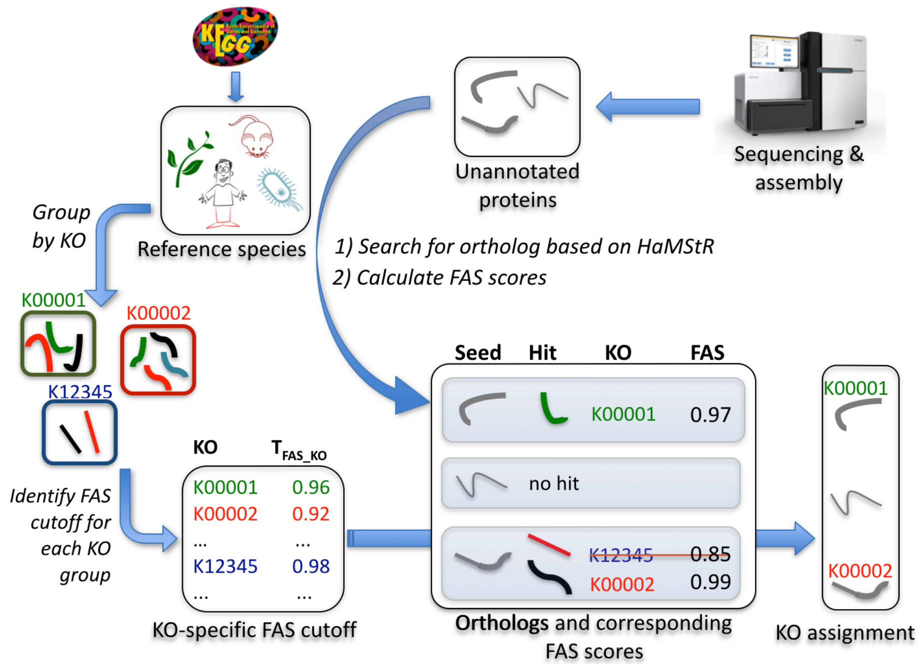

Figure 3-1: The workflow of a KO annotation transfer using HamFAS. See main text for a detailed description of the procedure.

The HamFAS approach can be distinguished into three main phases.

In phase one, the pre-processing phase, the protein sets of the 30 manually curated KO-annotated reference species provided by KEGG are retrieved. Sequences from different species sharing the same KO id are grouped, and we determine the all-vs-all pairwise FAS scores for each KO group using the python script `greedyFAS.py` implemented in HaMStR v13.2.9 (Ebersberger, Strauss, and von Haeseler 2009). The group's mean FAS score across all pair-wise comparisons determines then the KO specific FAS cutoff, T_{FAS_KO}.

In phase two, a targeted ortholog search of a query sequence in the 30 reference gene sets is performed with HaMStR v13.2.9, and the pair-wise FAS scores between the seed protein and its orthologs is computed using the python script `greedyFAS.py`. Each KO id that is represented in the identified orthologs is then a candidate KO id for the annotation of the seed protein.

In phase three, HamFAS analyzes, for each ortholog and for each candidate KO, if the observed FAS score between the seed protein and its ortholog falls within the diversity of FAS scores observed within the KO group (T_{FAS_KO}). If this is the case, the seed will be added to the KO group, and a functional annotation transfer is performed. Otherwise the KO group is not further considered in the annotation procedure. All KO ids that have FAS scores exceed the corresponding T_{FAS_KO} are used as valid annotations for the seed protein.

3.2.2 Materials and methods

List of 30 reference species download from KEGG database (http://www.genome.jp/kegg/) is described in Table 3-1 below.

Table 3-1: List of 30 manually KO-annotated reference taxa from KEGG.

Taxon ID	Taxon name	Phylum	Kingdom	Superkingdom
33169	*Ashbya gossypii*	Ascomycota	Fungi	Eukaryota
4896	*Schizosaccharomyces pombe*	Ascomycota	Fungi	Eukaryota
5476	*Candida albicans*	Ascomycota	Fungi	Eukaryota
4932	*Saccharomyces cerevisiae*	Ascomycota	Fungi	Eukaryota
5141	*Neurospora crassa*	Ascomycota	Fungi	Eukaryota
162425	*Aspergillus nidulans*	Ascomycota	Fungi	Eukaryota
9606	*Homo sapiens*	Chordata	Metazoa	Eukaryota

10090	*Mus musculus*	Chordata	Metazoa	Eukaryota
10116	*Rattus norvegicus*	Chordata	Metazoa	Eukaryota
7955	*Danio rerio*	Chordata	Metazoa	Eukaryota
7227	*Drosophila melanogaster*	Arthropoda	Metazoa	Eukaryota
6239	*Caenorhabditis elegans*	Nematoda	Metazoa	Eukaryota
81824	*Monosiga brevicollis*	Monosiga (genus)	NA$^{(*)}$	Eukaryota
45351	*Nematostella vectensis*	Cnidaria	Metazoa	Eukaryota
5759	*Entamoeba histolytica*	Entamoeba (genus)	NA$^{(*)}$	Eukaryota
5691	*Trypanosoma brucei*	Trypanosoma (genus)	NA$^{(*)}$	Eukaryota
3702	*Arabidopsis thaliana*	Streptophyta	Viridiplantae	Eukaryota
36329	*Plasmodium falciparum* 3D7	Apicomplexa	NA$^{(*)}$	Eukaryota
237895	*Cryptosporidium hominis*	Apicomplexa	NA$^{(*)}$	Eukaryota
2190	*Methanocaldococcus jannaschii*	Euryarchaeota	NA$^{(*)}$	Archaea
56636	*Aeropyrum pernix*	Crenarchaeota	NA$^{(*)}$	Archaea
511145	*Escherichia coli* str. K-12 substr. MG1655	*Proteobacteria*	NA$^{(*)}$	Bacteria
122586	*Neisseria meningitidis* MC58	*Proteobacteria*	NA$^{(*)}$	Bacteria
85962	*Helicobacter pylori* 26695	*Proteobacteria*	NA$^{(*)}$	Bacteria
224308	*Bacillus subtilis subsp.* subtilis 168	*Firmicutes*	NA$^{(*)}$	Bacteria
272623	*Lactococcus lactis subsp. lactis* Il1403	*Firmicutes*	NA$^{(*)}$	Bacteria
243273	*Mycoplasma genitalium* G37	*Tenericutes*	NA$^{(*)}$	Bacteria
83332	*Mycobacterium tuberculosis* H37Rv	*Actinobacteria*	NA$^{(*)}$	Bacteria
1148	*Synechocystis sp.* PCC 6803	*Cyanobacteria*	NA$^{(*)}$	Bacteria
63363	*Aquifex aeolicus*	*Aquificae*	NA$^{(*)}$	Bacteria

$^{(*)}$ *Undefined*

HaMStR v13.2.9 is downloaded from https://github.com/BIONF/HaMStR. In order to reduce the number of false positive orthologs, we made the orthology inference stricter by accepting only the reciprocal best hit from both HMM and reverse BLAST search within HaMStR procedure using the option `-rbh`.

FAS score calculation was done using `greedyFAS.py` script with the options `--priority_threshold 30` and `--weightcorrection loge` (correct the weight of protein features by log_e).

The KO transfer was performed using a custom Perl script `koAnnotation_hamfas.pl`.

3.3 Results and Discussion

3.3.1 The establishment of the reference species and annotations

We yielded in total 12,748 different KO groups from 30 KEGG reference species. The proteins in each group are very similar with each other in term of feature architecture, which can be accounted from the FAS score distribution in Figure 3-2.

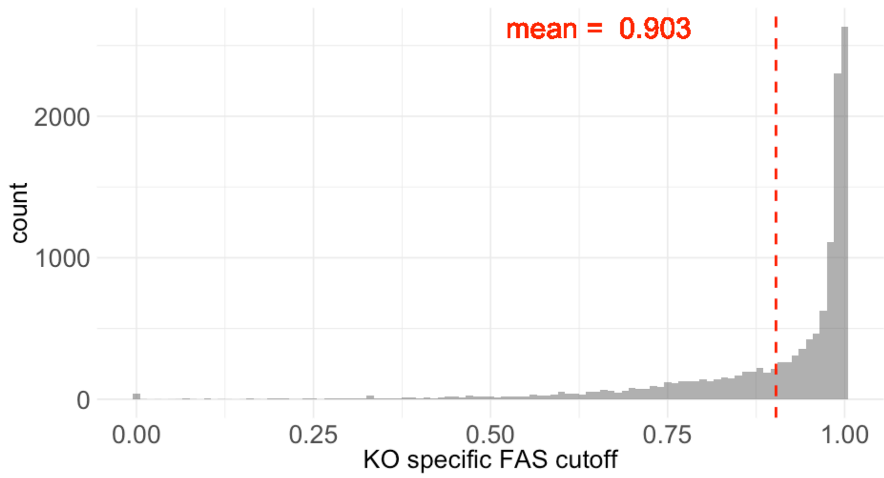

Figure 3-2: Distribution of T_{FAS_KO} for 12,748 KO groups.

Only about 3% of KO ids have T_{FAS_KO} smaller than 0.5, 27% lie between 0.5 and 0.9, while 70% has T_{FAS_KO} greater than 0.9. The low T_{FAS_KO} values are caused mostly by the poorly domain annotated protein members. Figure 3-3 shows 2 examples for representing a low T_{FAS_KO} group (K00542) and a high T_{FAS_KO} group (K0788).

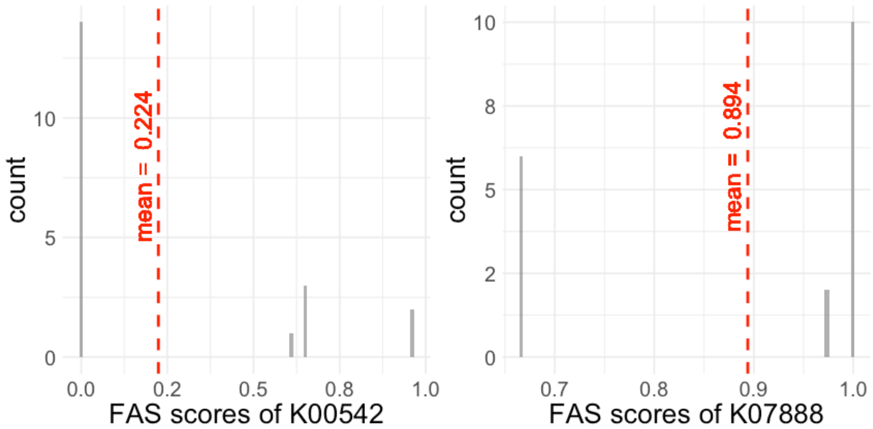

Figure 3-3: FAS score density of KO group K00542 (left) and K07888 (right).

In ortholog group K00542 (guanidinoacetate N-methyltransferase), only one protein member (rat rno:25257) has a single Pfam domain (Orn_DAP_Arg_deC). The lack of Pfam domain annotation of other proteins (human hsa:2593, mouse mmu:14431, zebrafish dre:796865 and *N.vectensis* nemve:1432) caused FAS scores of 0 for 14/20 pairwise comparisons and led to the low T_{FAS_KO} (mean score of 0.224) for the whole group. On the contrary, the rich annotation of protein members of group K07888 (Ras-related protein Rab-5B) is the reason for its high T_{FAS_KO}.

3.3.2 Benchmarking HamFAS

3.3.2.1 Outline of the benchmarking procedure

We used *Saccharomyces cerevisiae* (yeast) as a test species to benchmark the HamFAS approach. The general outline of the benchmark was as following: The protein set of yeast was obtained from KEGG database. It was divided into two subsets. Set 1 comprises the 3457 proteins that have an assigned KO id, and set 2 comprises the remaining 3158 proteins to which KEGG did not assign a KO id. To assess the specificity of the KO assignment by the three tools, we used the proteins in set 1, ignored their KO id, and reannotated them with HamFAS, BlastKOALA and KAAS. To compare the sensitivity of the three tools, we used the proteins in set 2, which do not carry a KO id, and annotated them with HamFAS, BlastKOALA and KAAS.

Because all three annotation tools, KAAS, BlastKOALA and HamFAS are based on homolog/ortholog searches in pre-annotated gene sets. To avoid circularity in the annotation process, we removed *S.cerevisiae* from the reference species used by both HamFAS and KAAS. In the case of BlastKOALA, which can be run only remotely on the servers of KEGG, a removal of the yeast proteins from the set of reference proteins was not possible. BlastKOALA uses a non-redundant data set compiled from the entire GENES database of KEGG (Kanehisa et al. 2016) for the KO assignment. No option is offered to customize this database.

3.3.2.2 The specificity of HamFAS

For assessing the specificity of the KO assignments by the various tools, we calculated the recall, precision and F1 score (equations 1, 2 and 3 respectively) for HamFAS and compared them with the F1 scores of BlastKOALA and KAAS, respectively.

$$recall = \frac{TP}{TP+FN} \ (1);\ \ precision = \frac{TP}{TP+FP} \ (2);\ F1 = \frac{2*precision*recall}{precision+recall} \ (3)$$

Table 3-2 summarizes the results of HamFAS, BlastKOALA and KAAS for set 1. Although the differences among the tools are subtle, HamFAS performed best in term of precision. However, its recall is only second to KAAS, which results in a slightly lower F1-score of HamFAS compared to KAAS. Without discarding S.cerevisiae sequences from the reference data set, the latest annotation tool from KEGG, BlastKOALA was expected to gain the best result. Interestingly, it yielded the lowest scores in both recall and precision.

Table 3-2: Recall, precision and F1-score of HamFAS in comparison to BlastKOALA and KAAS.

Approach	HamFAS	supported_HamFAS (*)	BlastKOALA	KAAS
Recall	0.915	0.861	0.905	0.931
Precision	0.985	0.985	0.979	0.984
F1-score	0.949	0.919	0.940	0.957

(*) *Result of HamFAS after filtering the orthology assignment with InParanoid's orthologs*

To see to what extent the stringency of the ortholog search affects the annotation transfer, we additionally used only such orthologs that are consistently identified by both HaMStR and InParanoid v4.1 (O'Brien, Remm, and Sonnhammer 2005). Note, InParanoid is among the ortholog search tools having the highest specificity maintaining at the same time a high sensitivity (Altenhoff et al. 2016). This additional filter left additional 188 yeast proteins un-annotated, when compared to the native HamFAS approach. It resulted in a decrease of the recall and F1-score, but did not affect the precision (see Table 3-2). Figure 3-4 shows that the FAS score distribution of the orthologs identified only by HaMStR, but not by InParanoid are slightly smaller than the ones of the orthologs supported by both ortholog search tools (Mann-Whitney-Wilcoxon's p-value $< 2.2e-16$). However, HamFAS decides on a transfer of KO ids using both the orthology assignment and a FAS score cutoff. Thus, a slightly higher inclusiveness in the ortholog search, even at the cost of including some possibly spurious orthologs will not affect the overall specificity of the transfer.

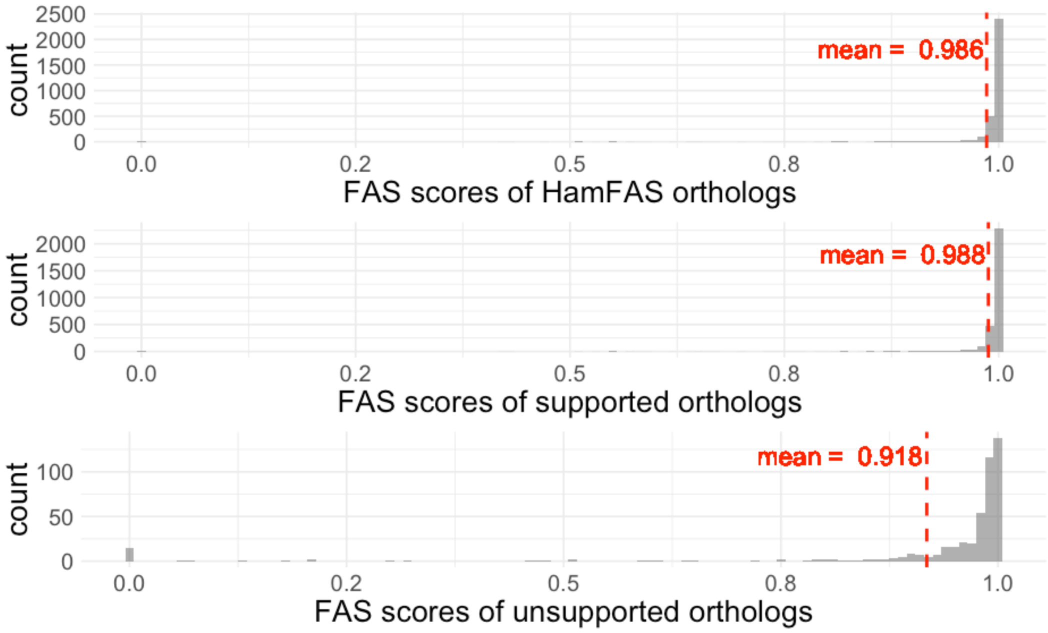

Figure 3-4: FAS score distribution of the orthologs detected in the course of the HamFAS analysis. The top histogram shows the distribution for all orthologs identified by HaMStR, the ortholog search tool natively implemented into HamFAS. The second histogram shows the FAS score distribution only for those orthologs that are consistently identified by both, HaMStR, and by InParanoid. Not the slight shift towards higher values. The third histogram displays the FAS score distribution for the orthologs only identified by HaMStR but not by InParanoid.

In the next step, we analyzed the fractions of proteins annotated by HamFAS, BlastKOALA and KAAS, respectively. The results are summarized in Figure 3-5. Most proteins (85.6%) are annotated by all 3 approaches, and only minor fractions are annotated only by one tool. In particular, 2.1% proteins were only annotated by BlastKOALA, and 0.6% in case of HamFAS or KAAS. In this context it might be interesting to note that BLastKOALA still contains yeast sequences in its reference set used for the annotation. This can explain that three times as many yeast proteins are annotated only by BLastKOALA, in comparison to the other two tools.

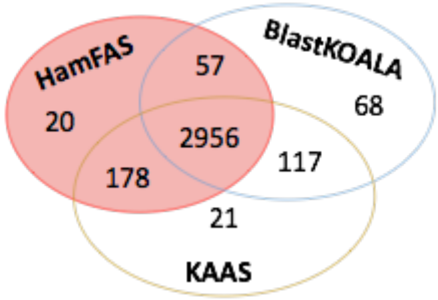

Figure 3-5: Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS. Note, at this level of the analysis, we checked only which proteins were annotated by which annotation tools. We did not check for the consistency of the annotation transfer.

For each protein that was annotated by at least two different approaches, we subsequently compared the transferred KEGG ids. It comes of little surprise that, in few cases, the same yeast protein is assigned different KO ids by the different tools (Table 3-3).

Taking a closer look at these deviating instances, most of them resolve to cases where the KO ids are “synonymous”. More precisely, we refer to two KO ids as "synonymous" if they either point to the same EC numbers, same EC classes, same GO numbers, or are the same components in KEGG pathways, responsible for the same reactions.

Table 3-3: Overview of the KO ids assigned to the yeast proteins in set 1 by HamFAS, BlastKOALA and KAAS.

Approach	All 3 approaches	HamFAS + BlastKOALA	HamFAS + KAAS	KAAS + BlastKOALA
Same KOs	2951	54	168	108
Diff. KOs	5 (1*)	3 (1*)	10 (5*)	9 (6*)
Total	2956	57	178	117

* *Different KO ids after filtered by synonymous KO ids.*

In summary, our evaluation based on the proteins in yeast set 1 has revealed that HamFAS is capable to reliably transfer functional annotation by means of KO ids to un-annotated proteins. Its performance is comparable to KAAS and slightly better than that of BlastKOALA.

3.3.2.3 The sensitivity of HamFAS

So far, our analysis was concentrating on the performance of the three tools in reproducing an already existing KO annotation of the yeast proteins. It is, thus, a considerably simplistic scenario. We next used the 3158 yeast proteins that have no KO id assigned by KEGG as input. For 257 of these, HamFAS assigned a KO id (Figure 3-6), of which 164 proteins are annotated only by HamFAS. In contrast, the KEGG tools KAAS and BlastKOALA assigned KO ids to only 150 and 116 proteins, respectively.

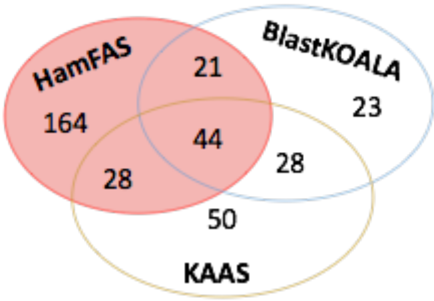

Figure 3-6: Fraction of proteins in the yeast set 2 for which HamFAS, BlastKOALA and KAAS each assigned a KO id. Only few proteins are annotated by each of the three tools. Most notably, the number of proteins that were annotated only by a single method is substantially higher of HamFAS, when compared to BlastKOALA and KAAS.

3.3.2.4 Analysis of annotations inferred by HamFAS

The results of the KO id annotation transfer for the yeast protein set 2 followed by and large the expectation. Only a small fraction of the total proteins in this set were annotated by at least one of the tools. This is not surprising given that the curation routines in KEGG did not assign these proteins a KO id in the first place. However, our analysis revealed that of the proteins that were annotated by HamFAS, more than half (164 / 257) are annotated only by our tool. This can either indicate a substantially higher sensitivity of HamFAS compared to KAAS and BlastKOALA. Alternatively, it can flag an elevated false positive rate. We therefore looked at the HamFAS predictions in closer detail.

We first asked whether the proteins annotated only by HamFAS (HamFAS-only proteins) differ in length when compared to the proteins that were additionally annotated by BlastKOALA or KAAS (control group). The results are shown in Figure 3-7. HamFAS-only proteins are neither significantly shorter nor longer than the proteins in the control group (Mann-Whitney-Wilcoxon's p-value = 0.78 and p-value = 0.38, respectively).

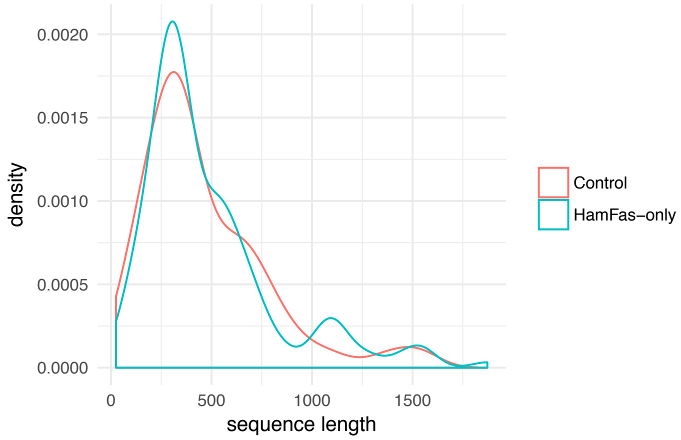

Figure 3-7: Length distribution of proteins in the HamFAS-only group and the control group (others).

Next, we asked whether there is any difference in the Pfam domain content between the proteins in the two sets (Figure 3-8). Again, no obvious difference was seen.

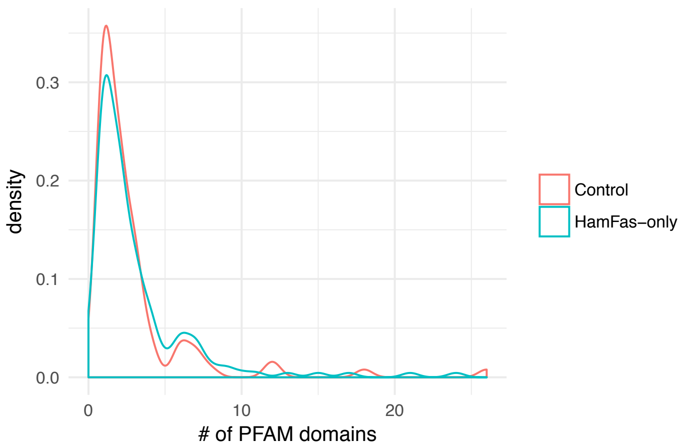

Figure 3-8: Distribution of the number of Pfam domains in the HamFAS-only proteins and in the proteins of the control group.

We asked further whether the feature architecture between the annotated yeast protein and its ortholog that served as a donor for the annotation transfer (the "annotation donor") are lower on average for the HamFAS-only group compared to the results (Figure 3-9). This revealed that the mean FAS score across all HamFAS-only proteins is with 0.936 slightly but statistically significantly smaller that the mean FAS score for the control group (0.947; Mann-Whitney-Wilcoxon's p-value = 0.001). Consulting the FAS score distribution in Figure 3-9, this difference appears to be driven by the higher number of proteins with a FAS score of 0 in the HamFAS-only set. These proteins lack any features, and thus the FAS score evaluates to 0. We therefore conclude, that there is no strong signal that the proteins annotated only by HamFAS deviate stronger in their feature architecture from the annotation donor than the proteins additionally annotated by KAAS or BlastKOALA.

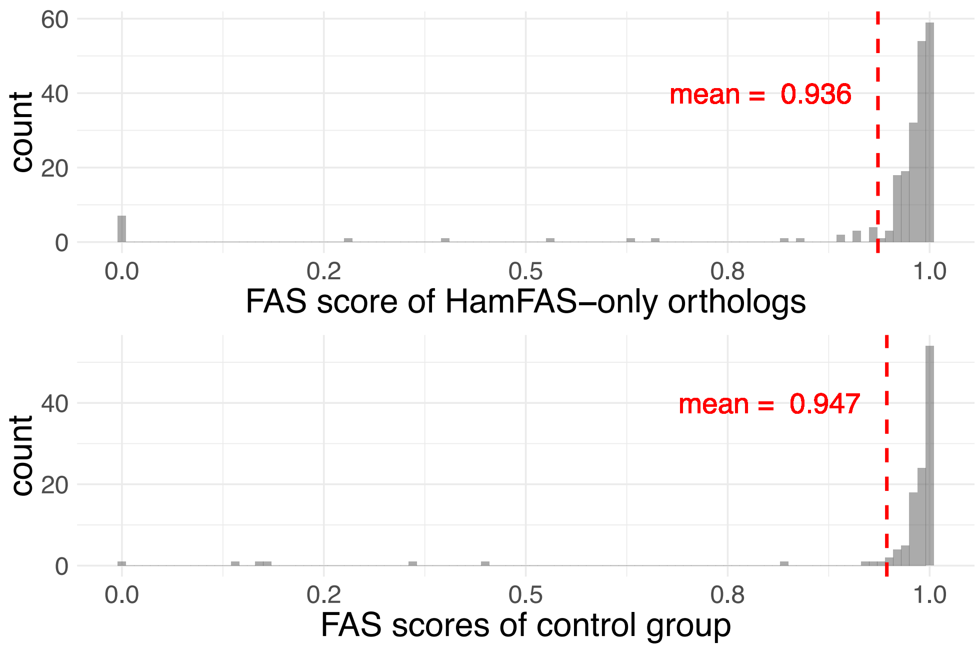

Figure 3-9: The distribution of FAS scores between the yeast proteins and their respective orthologs that served as donor for the annotation transfer. The histogram at the top represents the HamFAS-only proteins; the histogram at the bottom represents the score distribution for the proteins in the control group.

Next, we asked about the taxonomic distance between the yeast protein and the annotation donor. To this end, we grouped the annotation donors into the following taxonomic bins: Fungi, animals, other eukaryotes, archaea, and bacteria. We then plotted for each set, HamFAS-only, control, and additionally the set of yeast proteins that were already KEGG annotated (set 1 from 3.3.2.1 above), the fraction of annotation donors that were assigned to the five taxonomic bins (Figure 3-10).

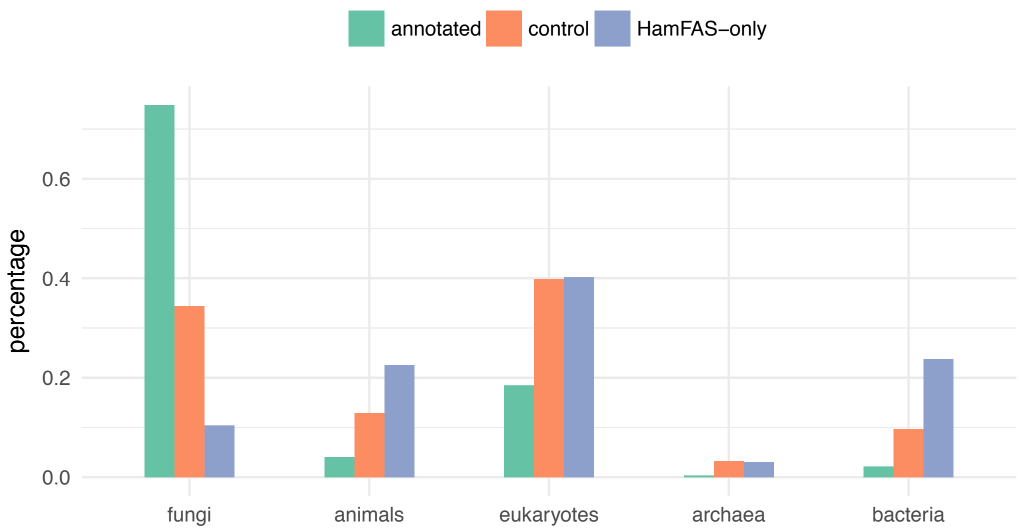

Figure 3-10: The fractions of annotations from fungi, animals, other eukaryotes, archaea or bacteria for KO-annotated, control and HamFAS-only protein set.

This revealed, not surprisingly, that for the already KEGG annotated yeast proteins, the annotation donor in our analysis was mostly (75%) a fungal sequence. The remaining annotation donors stem from either animals or from other eukaryotes, and only few of them have obtained annotations from archaea or bacterial taxa (2,4%). The situation is different for the proteins that are not already annotated by KEGG. Here, a considerable fraction of annotation donors stem from non-fungal orthologs, and from bacteria. This shift of the annotation donor towards more distantly related taxa is consistently seen for both the proteins in the HamFAS-only group and in the control group the latter. However, the control group appears to have a somewhat stronger preference for fungal annotation donors compared to the

HamFAS-only group. To shed further light on this issue, we subsequently excluded annotations from where the annotation donor was either of archaeal or bacterial origin. This did not affect the accuracy or the sensitivity of HamFAS (see Appendix, Table A-5 and Figure A-1). Furthermore, we analyzed the phylogenetic profiles of the yeast proteins in set 2 where the annotation donor was from a non-eukaryotic reference species. This analysis again revealed no noticeable difference between the HamFAS-only proteins and the proteins in the control group (see Appendix, Figure A-2 and Figure A-3).

In our last analysis, we considered protein interaction. First, we asked whether the connectivity of the newly functionally annotated proteins in set 2 differs between the HamFAS-only and the control group. To this end, we calculated the node degree for each yeast protein in the yeast protein-protein-interaction (PPI) networks retrieved from the Yeast Interactome Project (http://interactome.dfci.harvard.edu/S_cerevisiae/, (Yu et al. 2008)) and STRING database (https://string-db.org, (Szklarczyk et al. 2015)). We then plotted the distributions of the node degrees for the two categories in set 2, and again additionally for the yeast proteins that were already annotated by KEGG in set 1 (Figure 3-11).

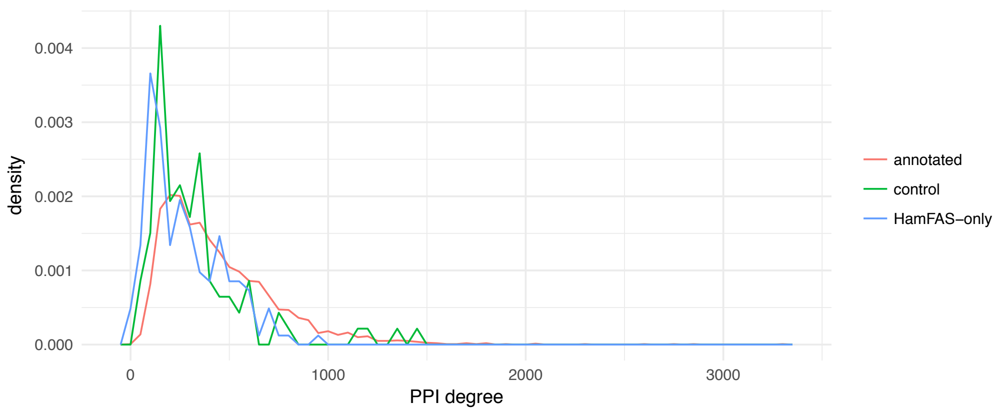

Figure 3-11: The PPI degree distribution of 3 protein sets.

The KO-annotated proteins have with mean PPI degree of 444 in general more interacting partners than the proteins in set 2. The proteins of the

HamFAS-only group have a mean PPI degree of 275, and the proteins in the control group have a mean PPI degree of 327. Notably, we did not observe any substantial difference of the PPI degrees for the HamFAS-only proteins, and those in the control set (p-value = 0.113).

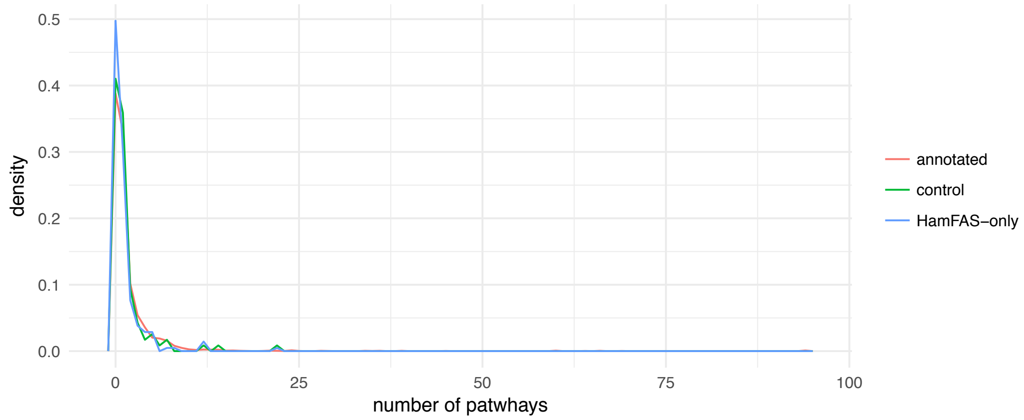

Figure 3-12: Distribution of the number of pathways in which annotated KO ids are involved.

Next, we inferred the number of pathways in which the KO ids are involved that have been assigned to the yeast proteins (Figure 3-12). Overall, more than half of the KO ids are represented in at least one KEGG pathway. Again, the percentage is highest for the KO ids that have been assigned to the pre-annotated proteins in set 1 (61%). For the proteins in set 2, the values are lower and similar for the HamFAS-only set and the control group (50% and 58 % respectively).

In summary, our analyses so far have revealed that the yeast sequences annotated only by HamFAS do not differ substantially from the sequences that are annotated also at least by one of the other tools. There is only one notable exception. It appears that the taxonomic distance of the annotation donor seems to differ between the proteins in the HamFAS-only set and the proteins in the control group, where the latter have a tendency to be annotated by evolutionarily more closely related sequences. Overall, our analyses have provided no indication that the considerably high number of proteins annotated only by HamFAS is due to a reduced specificity of our tool. Instead, it appears that it has, indeed, a higher sensitivity when compared to BlastKOALA and KAAS. In line with this notion, we find that the additional annotations by HamFAS can complement a total of 29 pathways of which components have been already identified in the yeast gene set. See Figure 3-14 and further examples in the Appendix, Figure A-4, Figure A-5 and Figure A-6.

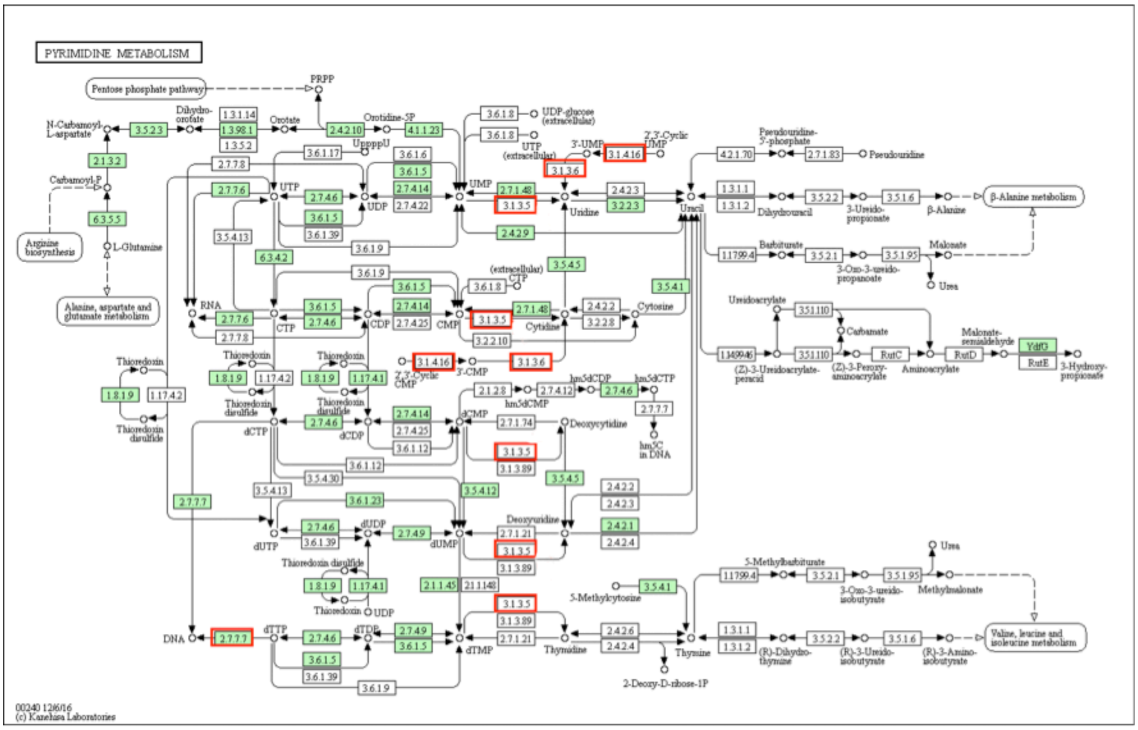

Figure 3-13: Pyrimidine metabolism for HamFAS annotated yeast proteins. Green highlighted boxes are yeast proteins already present in the KEGG database. Red boxes are complementary proteins from the HamFAS-only annotation. The pathway scheme was obtained from KEGG.

The total set of pathways to which the KO ids that have been assigned by HamFAS only is shown in Figure 3-14.

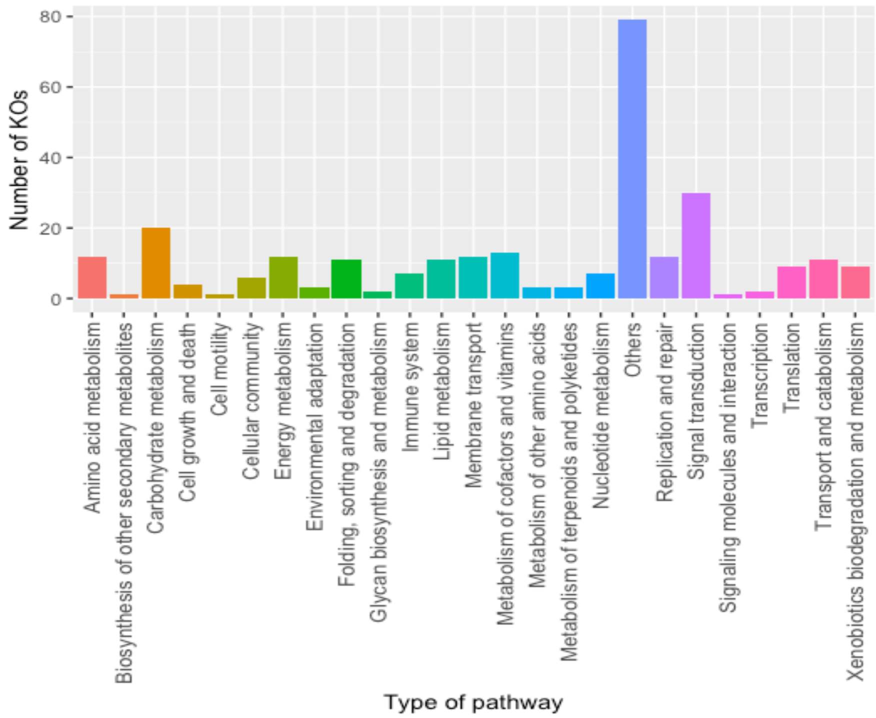

Figure 3-14: The numbers of HamFAS-only KO ids assigned to different pathway categories.

3.4 Conclusion

Our analyses exemplarily performed on the yeast protein set have shown that HamFAS is comparable in specificity to the two state of the art tools, BlastKOALA and KAAS, and appears to have a slightly higher sensitivity. This demonstrates that the combined use of an ortholog search tool, HaMStR, and of a measure of the pairwise feature architecture similarity between proteins, represent a powerful approach for a functional annotation transfer. The conceptual simplicity of HamFAS, in contrast to the other two tools – and in particular to BlastKOALA – is appealing. In essence, it allows to use HamFAS without any modification in combination with any group of functionally equivalent proteins provided from resources other than KEGG. HamFAS is straightforward to install and can be run on a local computer architecture. This allows a full control of the protein annotation procedure. Specifically, users can choose different methods for the orthology assignment step, and they can alter the FAS filtering thresholds to increase or reduce the stringency of the annotation procedure. Last but not least, HamFAS removes the dependency on remote computer architecture together with all the connected constraints. By that it improves on BlastKOALA which can be run only remotely on the servers provided by KEGG, and for which it appeared impossible to adjust the reference gene set used during the annotation transfer. Eventually, BlastKOALA is limited to 5000 queries per job submission, and thus requires extensive manual interaction if larger gene sets should be annotated. Taken together, HamFAS is a novel and versatile software that complements the still considerably small collection tools for a functional annotation transfer.

4 Tracing the evolution of the microsporidian gene set

4.1 Introduction

Microsporidia are considerably simple organized organisms that share a common ancestry with the fungi (Hibbett et al. 2007). Probably as an adaptation to their lifestyle as an obligate intracellular lifestyle, microsporidia have substantially reduced both their genomes and their gene sets (Agnew et al. 2003; Williams 2009; Nakjang et al. 2013). However, the evolutionary trajectory of this reduction process is still not entirely understood. In particular, one can raise the question whether the last common ancestor (LCA) of the microsporidia was already as reduced as the contemporary species. In other words, was this LCA most likely already an obligate intracellular parasite? Or did the reduction process and the change of the lifestyle happen several times independently within the microsporidia. Using the tools described in the previous chapters, we performed a comparative gene set analysis of the microsporidia to address these questions.

4.1.1 Phylogenetic tree and the last common ancestor

The evolutionary relationship between genes or species is typically represented as a phylogenetic tree (Figure 4-1). Originally, phylogenetic trees were used solely for the systematic classification of species (Choudhuri 2014). Nowadays, phylogenetic trees are regularly used in, and provide the fundamental backbone of comparative gene set analyses. They are prerequisites for tracing the evolutionary history of genes and of their functions across species and through time (Soltis and Soltis 2003; Gabaldón 2007; Gaucher, Kratzer, and Randall 2010).

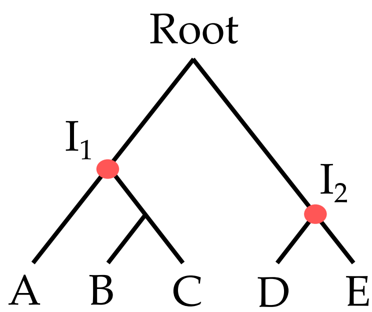

Figure 4-1: A schematic species tree demonstrates the phylogeny of five species A, B, C, D and E. Each leaf represents an individual species, while an internal node represents the last common ancestor of all leaf taxa that are connected to this node. For instance, I_1 is the last common ancestor of A, B and C. Similarly, I_2 is the last common ancestor of D and E.

In a species tree, the contemporary species are placed at the leafs of the tree, and the internal nodes denotes hypothetical ancestral taxa. In particular, the internal node that is closest to a set of contemporary species is denoted the last common ancestor (LCA) of the species connected to this node (Gregory 2008). The LCA could be a single organism or a population of cells that contained most of the shared features and encoded genes between its descendants (Moreira and López-García 2007). This common ancestry, the core of evolutionary theory, provides a basic for understanding the evolution of the contemporary species (Baum, Smith, and Donovan 2005).

4.1.2 The role of the microsporidian LCA in the understanding of their evolution

The analysis of phylogenies can give insight into the evolutionary history of species, such as what is the systematic relationship between species in the phylogenetic tree of life, or how their pathways evolved across taxa and time (Futuyma 2005). Since the evolutionary process of microsporidia is still poorly understood, a comparative analysis between the contemporary microsporidia and their ancestor is required (Keeling and Fast 2002). Investigation of the last common ancestor of microsporidia can give insight to many aspects of their evolutionary history. Such as, the compact genomes of the extant microsporidia were the ancestral state or the result of a reduction process, the different fraction between microsporidian gene sets was the result of the losses from the ancestral gene set or gain events in the individual species, or what the ancient metabolism was and how it related to the parasitic lifestyle of microsporidia.

For that reason, in this chapter we describe an orthology-based approach for estimating the microsporidian LCA protein set, which was served as an initial data for further analyses. Furthermore, we investigated the evolutionary history of microsporidian proteins and confirmed the fungal-relationship of microsporidia based on their phylogeny. Lastly, we reconstructed the metabolic pathways of the microsporidian LCA and compared them with the current knowledge about the extant microsporidian metabolism to gain insight about the origin of their obligate intracellular parasitic lifestyle.

4.2 Methods

4.2.1 Data

Microsporidia taxa

For this study, we compiled a data set comprising 11 microsporidian species whose genomes have been fully sequenced. The data was obtained from the following resources: The genome portal of the JGI database of Join Genome Institute (Nordberg et al. 2014) and the MicrosporidiaDB (Aurrecoechea et al. 2011) of the microsporidia genome sequencing project of the Broad Institute (Cuomo et al. 2012; Pombert et al. 2013; Bakowski et al. 2014; Desjardins et al. 2015). The species name, strain name, number of proteins as well as the source database of those eleven microsporidia can be found in Table 4-1.

Table 4-1: Taxon set A - The microsporidia data set that used in this project.

Name	Strain	Number of proteins	Source
Encephalitozoon hellem	ATCC 50504	1827	JGI
Encephalitozoon intestinalis	ATCC 50506	1657	Broad Inst
Encephalitozoon cuniculi	GB-M1	1896	Broad Inst
Nosema ceranae	BRL01	2057	Broad Inst

Enterocytozoon bieneusi	H348	3312	JGI
Vittaforma corneae	ATCC 50505	2243	Broad Inst
Anncaliia algerae	PRA339	3576	Broad Inst
Antonospora locustae	HM-2013	2191	JGI
Edhazardia aedis	USNM 41457	4208	Broad Inst
Vavraia culicis subsp. floridensis	-	2775	Broad Inst
Nematocida parisii	ERTm1	2659	Broad Inst

Taxa for microsporidian LCA protein set reconstruction

For inferring the gene set of the microsporidian LCA, we additionally selected 17 opisthokonts comprising 13 fungi, 2 animals, *Monosiga brevicollis* and *Capsaspora owczarzaki.* This collection of opisthokont species resembles the taxon set that was used in the study of (Capella-Gutiérrez, Marcet-Houben, and Gabaldón 2012) who aimed at elucidating the phylogenetic position of the microsporidia. We further complemented that taxon set with 7 bikonts to serve as an outgroup. We refer to this taxon collection as Taxon set B, and provide further details in Table 4-2.

Table 4-2: Taxon set B - 24 taxa used for reconstructing the microsporidian LCA protein set.

Taxon ID	Taxon name	Phylum	Kingdom	Source
4932	*Saccharomyces cerevisiae*	Ascomycota	Fungi	Ensembl (1)
5476	*Candida albicans*	Ascomycota	Fungi	CGD (2)
5141	*Neurospora crassa*	Ascomycota	Fungi	UniProt (3)
162425	*Aspergillus nidulans*	Ascomycota	Fungi	Broad Inst (4)
4896	*Schizosaccharomyces pombe*	Ascomycota	Fungi	UniProt (3)
29883	*Laccaria bicolor*	Basidiomycota	Fungi	JGI (5)
5297	*Puccinia graminis*	Basidiomycota	Fungi	Broad Inst (4)
36080	*Mucor circinelloides*	Mucoromycota	Fungi	JGI (5)
64495	*Rhizopus oryzae*	Mucoromycota	Fungi	Broad Inst (4)
4837	*Phycomyces blakesleeanus*	Mucoromycota	Fungi	JGI (5)
109871	*Batrachochytrium dendrobatidis*	Chytridiomycota	Fungi	JGI (5)
109760	*Spizellomyces punctatus*	Chytridiomycota	Fungi	Broad Inst (4)

281847	*Rozella allomycis*	Cryptomycota	Fungi	UniProt [3]
45351	*Nematostella vectensis*	Cnidaria	Metazoa	JGI [5]
400682	*Amphimedon queenslandica*	Porifera	Metazoa	UniProt [3]
81824	*Monosiga brevicollis*	NA	NA	JGI [5]
192875	*Capsaspora owczarzaki*	NA	NA	Broad Inst [4]
5833[*]	*Plasmodium falciparum*	Apicomplexa	NA	plasmoDB [6]
237895[*]	*Cryptosporidium hominis*	Apicomplexa	NA	NCBI [7]
5691[*]	*Trypanosoma brucei*	NA	NA	Sanger Inst [8]
5762[*]	*Naegleria gruberi*	NA	NA	JGI [5]
3702[*]	*Arabidopsis thaliana*	Streptophyta	Viridiplantae	UniProt [3]
3055[*]	*Chlamydomonas reinhardtii*	Chlorophyta	Viridiplantae	JGI [5]
67593[*]	*Phytophthora sojae*	NA	NA	JGI [5]

(*) *Outgroup taxa*

(1) *Ensembl (https://www.ensembl.org/index.html)*

(2) *Candida Genome Database (CGD, http://www.candidagenome.org)*

(3) *UniProt (http://www.uniprot.org)*

(4) *Broad Institute (https://www.broadinstitute.org)*

(5) *Join Genome Institute (JGI, https://jgi.doe.gov)*

(6) *The Plasmodium Genomics Resource (PlasmoDB, http://plasmodb.org/plasmo/)*

(7) *The National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov)*

(8) *Sanger Institute (http://www.sanger.ac.uk/science/data)*

Taxa for fungal tree reconstruction

This taxon set C comprises 48 fungi representing Ascomycota, Basidiomycota, Blastocladiomycota, Chytridiomycota, Entomophthoromycota, Glomeromycota, Neocallimastigomycota, Kickxellales, Mortierellales and Mucorales, together with 11 microsporidia in Table 4-1, 6 opisthokonts including 3 animals, *Monosiga brevicollis*, *Capsaspora owczarzaki*, and *Fonticula* alba, and 7 bikonts as outgroup in Table 4-2. The list of taxon names, their NCBI taxonomy identifiers, systematic ranks and sources for their proteomes are provided in Table A-2 in the Appendix.

Taxa for analysis of microsporidian LCA's phylogenetic profiles

We used the gene sets of 491 species distributed across the tree of life including eukaryote, archaea and bacteria (Appendix, Table A-1) to perform a comprehensive analysis of the phylogenetic distribution of the microsporidian LCA proteins. These 491 species represent 44 higher order systematic groups and represent the Taxon set D (Figure 4-2).

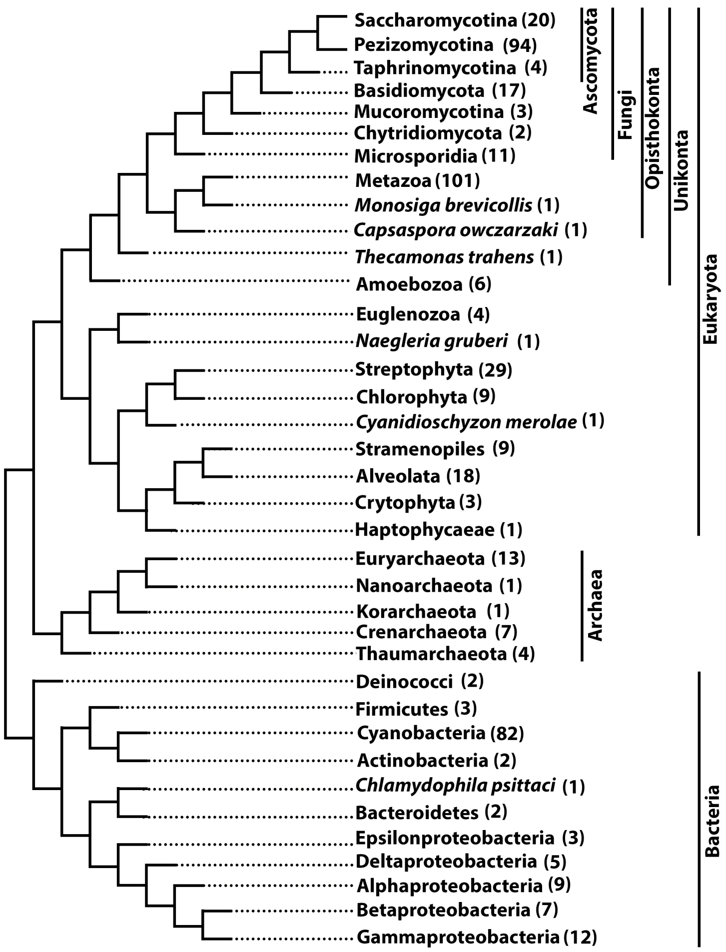

Figure 4-2: Tree representation of Taxon set D. The tree comprises all three domain of life (Woese, Kandler, and Wheelis (1990)) and displays all higher order taxa that are represented in Taxon set D. The number of species subsumed under each taxon is given in parenthesis. The tree is adapted from Ebersberger et al. (2014).

Reference taxa for the KEGG Orthology annotation

We annotated the microsporidian proteins with KEGG KO ids using the HamFAS approach described in Chapter 3. As described in 3.2.2, we used 30 manually curated KO annotated species downloaded from KEGG database as reference species for our KO annotation pipeline. NCBI IDs, taxon names and taxonomy ranks of those species are given in Table 3-1.

4.2.2 Orthologs search

Orthology assessment is one of the most crucial steps in the course of performing a comparative gene set analysis (Gabaldón 2008). Several automatic orthology prediction methods based on sequence similarity and / or phylogenetic tree have been developed (Kristensen et al. 2011). Although a phylogenetic tree can explain best the homologous relationship between genes, the tree-based approaches are can become highly time-consuming depending on the complexity of the data. In contrast, the sequence similarity-based methods, can be applied for a large amount of data in an affordable time (Trachana et al. 2011). In our study, we used the latter approaches for the orthology predictions.

Searching orthologs within microsporidia using OrthoMCL

We used OrthoMCL v2.0.9 (Li, Stoeckert, and Roos 2003) to search for orthologous proteins within the selected microsporidia taxon set. OrthoMCL performs an all-against-all BLASTP (Altschul et al. 1997) comparison for all input taxa and clusters homologous groups together using the Markov Cluster algorithm MCL (van Dongen 2000). OrthoMCL was run with standard parameters i.e. `-v 100000 -b 100000 -z 0 -e 1e-5` for the BLAST search and an MCL inflation parameter of 1.5.

Extending orthologous group using HaMStR

We extended the microsporidian orthologous groups retrieved from OrthoMCL by searching for their orthologs in the other taxa (cf. Table 4-2, or Table A-1 and Table A-2 in the Appendix) with HaMStR v13.2.9 from https://github.com/BIONF/HaMStR (Ebersberger, Strauss, and von Haeseler 2009). To this end, we used each orthologous group predicted by OrthoMCL as training data for a corresponding profile hidden Markov model (HMM) (Eddy 1998). HaMStR then used these HMM profiles in a targeted search to identify orthologs in further species outside the microsporidia. Each candidate protein obtained by the HMM search were added into the original orthologous group, if it fulfilled the reverse BLAST search (Altschul et al. 1990) criteria. In the reverse BLAST search, the candidate protein was searched against the proteomes of the seed species in the original orthologous group. By default, the candidate protein will be confirmed as a new ortholog, if the best hit from the reverse BLAST search is the same as the seed sequence. As microsporidia genes tend to evolve quickly (Slamovits et al. 2004), the BLAST search could be false to return the seed sequence as its best hit. We therefore run HaMStR with the options *`-checkCoorthologsRef`* to increase the sensitivity of the prediction by accepting the seed protein to be co-orthologous to the best reverse BLAST hit. Besides, we used other options to increase the specificity of HaMStR, including *`-hit_limit = 10`* to take only the first ten hits from the HMM search, *`-strict`* to force the candidate protein has to be orthologous with all seed proteins in the original group, and *`-representative`* to select only one ortholog for each search species.

4.2.3 Phylogenomic tree reconstruction

To reconstruct the evolutionary relationships of the microsporidia (cf. taxon set A in Table 4-1) and of the other taxa in our study (cf. Table 4-2, or Table A-2 in the Appendix), we pursued a standard phylogenomic approach using a supermatrix approach (Kupczok, Schmidt, and von Haeseler 2010). First, we identified a set of core genes, i.e. of genes that are represented as one-to-one orthologs in each of the taxa in the taxon set A and B in Table 4-1 and Table 4-2 (for example gene D in Figure 4-3). The protein sequences for each of the corresponding orthologous groups were then aligned with ClustalW v2.1 (Larkin et al. 2007). Subsequently, we concatenated the corresponding alignments into a supermatrix using a custom Perl script. To eliminate data that contain poor phylogenetic signals, we removed alignment columns with more than 50% of gaps. Subsequently, we selected the model to be used in the maximum likelihood tree reconstruction with ProtTest v3.4 (Abascal, Zardoya, and Posada 2005) using the post-processed supermatrix. Based on the best model parameters obtained from ProtTest, we reconstructed 100 bootstrap trees from the post-processed supermatrix with RAxML v8.1.9 (Stamatakis 2014). We increased the value of the seeds (parameter `-p` and `-b`) from 5 to 500 by a stepwise of 5. The consensus tree from those 100 individual maximum likelihood trees was then created by TREE-PUZZLE v5.3.rc16 (Schmidt et al. 2003). Lastly, we added the bootstrap supported values into the consensus tree with RAxML v8.1.9 using the parameter `-f b`. The final tree was then rooted using the taxa outside of the opisthokonts as outgroup.

4.2.4 Analysis of the microsporidian pan-gene set

We characterized the orthologous proteins that are shared between microsporidian species and in addition also those proteins that occur only in a single species in our data set. We refer to the latter proteins as orphans. First, we compare the length distributions of those two gene categories with the nonparametric U-test Wilcoxon-Mann-Whitney (Mann and Whitney 1947). Then, we performed a protein family (Pfam) domain annotation analysis for the orphan and orthologous proteins in each microsporidian species. To this end, we use hmmscan from the HMMER package v3.1b2 (Finn et al. 2015) in combination with the profile hidden Markov models from the Pfam-A database (Finn et al. 2016). Pfam domains represent, in general, evolutionarily conserved sub-sequences in a protein, of which a considerable fraction has been associated with a particular function (Finn et al. 2014).

4.2.5 Reconstruction of the microsporidian LCA gene set

For estimating the microsporidian LCA proteins, we performed a two stage analysis. First, we used OrthoMCL to search for orthologs within the collection of microsporidian species (cf. Table 4-1). In the second step, we used the OrthoMCL orthologous groups as so called 'core orthologs' in a HaMStR search. We used these core orthologs to train the corresponding profile hidden Markov models and then used HaMStR v13.2.9 to extend the OrthoMCL core ortholog groups with sequences from further species shown in Table 4-2.

Using the principle of minimum evolution (Edwards 1996), we filtered the orthologous groups based on the reconstructed maximum likelihood tree to identify the final protein set represented in the microsporidian LCA. The general procedure was as following: We projected each orthologous group onto the species tree and approximated the evolutionary age of the microsporidian proteins using a last common ancestor (LCA) approach (Capra et al. 2013). In essence, we identified for each orthologous group the two most distantly related species in this group. Their last common ancestor served then as an age estimation for the microsporidian proteins. To assign a microsporidian protein to the LCA set, we required at least one of the two following conditions to be met. (1) a protein must be represented by an ortholog in the earliest branching microsporidian lineage plus at least in one other microsporidian lineage (gene A and gene B in Figure 4-3). (2) a protein must be represented by at least two orthologs within the microsporidia and additionally in one or more species outside the microsporidia (gene C in Figure 4-3). The LCA set inference was done with the custom Perl script `filterList_2micros.pl`.

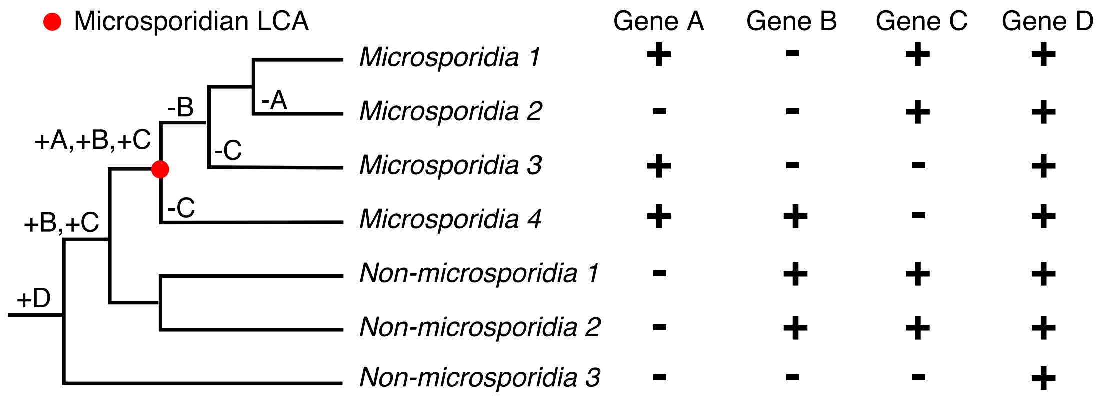

Figure 4-3: Different evolutionary scenarios of microsporidian LCA genes. Plus (+) and minus (-) represent the presence and absence of a gene in a species, respectively. Gene A represents the set of genes that are present only in the microsporidian lineage. B is a gene that exists in taxa outside of the microsporidia and in the earliest branching microsporidia species (*Microsporidia 4*). Gene C was present before the event that split microsporidia from the other taxa but it got subsequently lost on two microsporidian lineages leading to the taxa *Microsporidia 3, and 4*. Gene D is present in all taxa. It exemplifies the situation for a gene that is added to the set of core genes used for reconstructing phylogenetic trees.

4.2.6 Phylogeny of fungi

We expanded the phylogenetic analysis in the previous step to a comprehensive set of fungal species (taxon set C, Table A-2 in Appendix) in order to investigate the relationship of microsporidian and fungi. As a data basis for this analysis, we selected the one-to-one orthologs that are present in each of the species analyzed in 4.2.3, i.e. the set of core genes. Then, we performed an ortholog search with HaMStR v13.2.9 for those microsporidian core gene set in the taxon set shown in Table A-2. On the basis of these results, we then reconstructed the phylogenetic tree based on the procedure described in 4.2.3.

We additionally compared the acquired tree topology with existing hypotheses about microsporidian origins to see whether our tree explains the data significantly better. We used Beth (an in-house software developed by Ben Haladik, 2018) to reorder the taxa in our tree according to the hypotheses discussed in 1.2.3 of Chapter 1. We then used CONSEL v0.20 (Shimodaira and Hasegawa 2001) to test whether the alternative tree topology can be rejected as it provides a significantly worse hit compared to the maximum likelihood tree. Those tests include the approximately unbiased test (Shimodaira 2002), bootstrap probability of the selection (Felsenstein 1985), Bayesian posterior probability (Rannala and Yang 1996), Kishino-Hasegawa test (Kishino and Hasegawa 1989), Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999), weighted Kishino-Hasegawa test and weighted Shimodaira-Hasegawa test. Two tree topologies are significant different, if the P-values from the tests are less than 5%.

4.2.7 Phylogenetic profile analysis

We analyzed the phylogenetic profiles of the microsporidian LCA proteins by using HaMStR v13.2.9 to search for their orthologs in 491 taxa listed in taxon set D (cf. Table A-1 in Appendix). HaMStR was run with the same options as described in 4.2.2. To further complement the orthology assignment, we calculated the feature architecture similarity scores (Koestler, von Haeseler, and Ebersberger 2010) (FAS scores) for all pairwise proteins between the microsporidian seed protein and its orthologs using the python script `greedyFAS.py` implemented in HaMStR v13.2.9 with the options `--priority_threshold 30` and `--weightcorrection loge`.
At the end, we applied PhyloProfile (Tran, Greshake Tzovaras, and Ebersberger 2018) to analyze the phylogenetic profiles of the microsporidia LCA proteins with FAS scores as the complementary information to the presence/absence of the orthologs across 491 selected taxa.

4.2.8 Functional annotation and metabolic pathway mapping

KEGG Orthology assignment using HamFAS

We applied HamFAS approach to perform KO annotation for the microsporidian LCA proteins. Because one microsporidian LCA protein is represented by an orthologous group of several members, we assigned the representative FAS score for each reference protein as the max score that protein can achieve when compare with all microsporidia proteins in the corresponding orthologous group. The K numbers of reference proteins, which have the representative FAS score exceeded the T_{FAS_KO}, were transferred to that microsporidian LCA protein.

Besides complementing FAS scores to the orthology assignment, we also measured the patristic distance (Fourment and Gibbs 2006) between the reference protein and microsporidia protein to use it as a confidence value for the annotation transfer. The patristic distances were calculated from the reconstructed gene tree based on RAxML using the Python DendroPy library (Sukumaran and Holder 2010). The distance of a reference ortholog i in the orthologous group G is normalized to a range of [0,1] by the formula (2):

$$normalized\_dist(i) = \frac{dist(i) - min\_dist(G)}{max\_dist(G) - min\_dist(G)} \quad (2)$$

in which, $min\_dist(G)$ and $max\_dist(G)$ is minimal and maximal distance between that reference ortholog i to all microsporidia proteins in the orthologous group G.

Finally, we chose the confidence value for each annotated KO id as the lowest normalized distance among all reference proteins that have the matching KO id.

KEGG pathway mapping

To gain knowledge about the metabolism of the microsporidian LCA, their KO-annotated proteins were analyzed within the KEGG pathways. Those mapped pathways were further compared with the one of *E.cuniculi*, *E.hellem*, *E.intestinalis*, *N.ceranae*, four out of eleven contemporary microsporidia species under this study that are available in KEGG database, together with *S.cerevisiae* as an example for the free-living organism. The annotations and pathway information for those extant species were retrieved directly from KEGG database.

First, we analyzed the connectivity of microsporidian LCA and the contemporary species to gain the impression about their distribution in the metabolic network. For each reference KEGG pathway, the connectivity network nodes are enzymes (represented by their KO ids) in the pathway and edges are links between those nodes.

Then, we mapped the KO annotated proteins into the KEGG reference pathways for a more detailed investigation.

4.3 Results

4.3.1 The evolutionary history of the microsporidian pan-gene set

We used the gene sets of eleven microsporidian species as a starting point of this analysis. As a first step, we classified the evolutionary relationships of these proteins using OrthoMCL (Li, Stoeckert, and Roos 2003). The 20,485 proteins were grouped into 2904 orthologous comprising between 2 and maximally 148 proteins.

Based on the results of the ortholog search, we investigated the evolutionary history of the genes encoded in the contemporary microsporidian genomes. As a start, we distinguished two fractions, those genes with at least one ortholog in another species, and those genes for which OrthoMCL could not detected an orthologous protein in the other microsporidian species. In the following, we refer to these latter genes as *"orphans"*. When focusing on the individual species in our microsporidian set, we noted a considerable variation in the fraction of orphans (Figure 4-4). Most notably, the number of orphans in most cases increases with the total number of genes annotated in a genome. The three species from the genus *Encephalitozoon* have with 27 – 40 the fewest orphans among all microsporidia analyzed. At the same time, they have the smallest genomes and the smallest gene sets among all microsporidia considered here. In these species, only 2% of the genes appear as orphans. In turn, orphans make up about 1/5th of the annotated genes in *N. ceranae*, and almost half (49%) of the genes in the genome of *Edhazardia aedis*. Note, the genome of *E. aedis* is, with a length of over 50 Mbp, about an order of magnitude larger than that of *Encephalitozoon*.

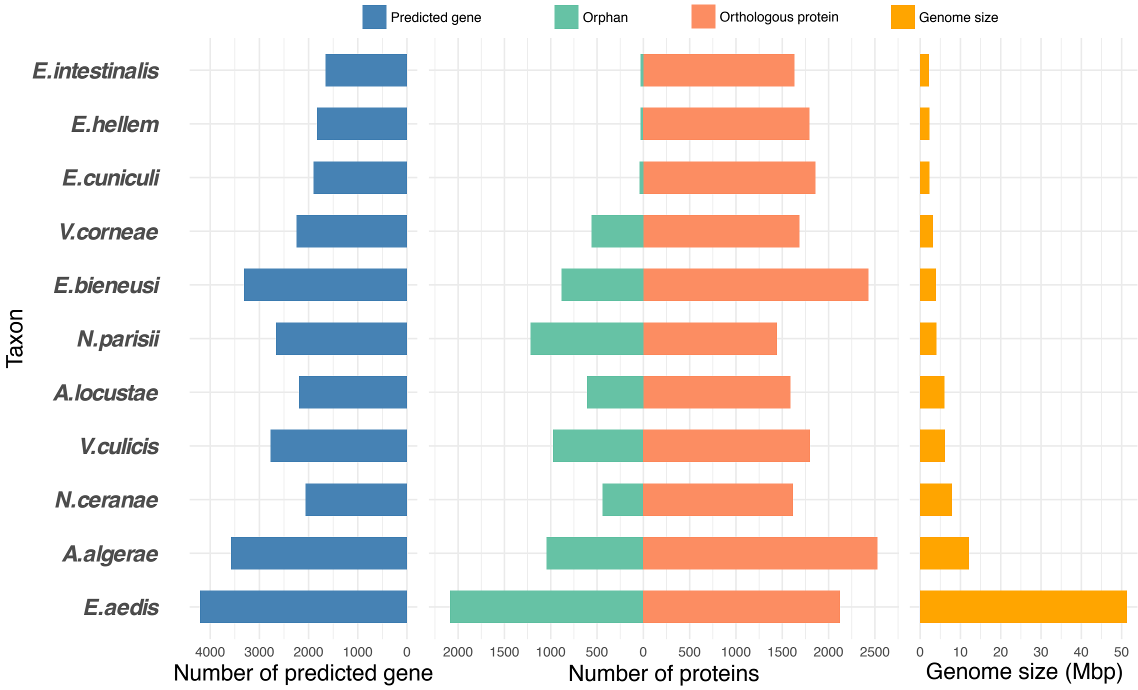

Figure 4-4: Number of total predicted genes (blue), orphan (green) and orthologous proteins (orange) in eleven microsporidian species. Taxa are ordered from top to bottom by increasing genome size (yellow).

Subsequently, we characterized the two gene categories in greater detail. First, we compared for each species the length distributions of orphans to that of genes with orthologs in other species. Figure 4-5 shows that orphans are, with an average length of 267 (exemplarily for *A.locustae* 158, *E.bieneusi* 182, *N.ceranae* 279, or *N.parisii* 302), about 72 amino acids shorter than genes with orthologs ($p<0.05$). The sole exception is *E. hellem*, where the length difference is not significant. However, this species harbors only 32 orphans, suggesting that the small sample size might interfere with the power of the test (Noether 1987).

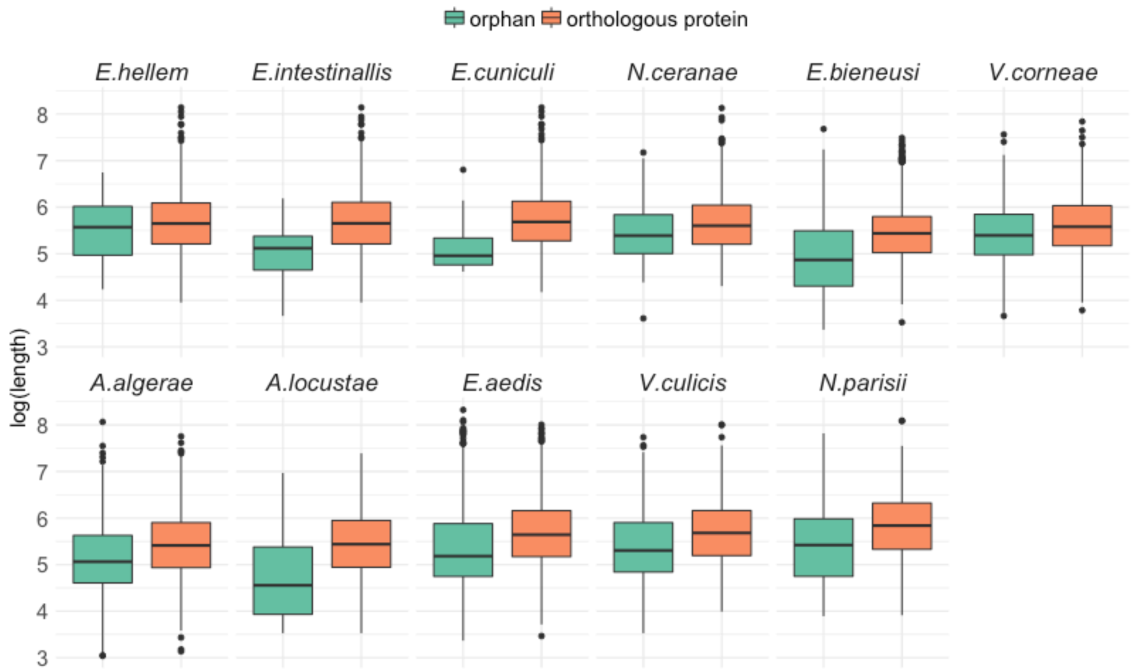

Figure 4-5: Length distribution of orphan proteins (green) and proteins with orthologs (orange) in the microsporidia. The length on the y axis is given in log(e) scale.

In the next step, we determined – again for each species separately - the presence of Pfam domains in the two gene sets. The results of this analysis are summarized in Figure 4-6. Our search revealed that the majority of genes with orthologs in other species do harbor at least one Pfam domain, and only between 24% and 39% of these genes are devoid of any Pfam domain. The situation is reversed for the orphans. Here, the majority of the proteins do not contain a Pfam domain, and only between 8 (microsporidium *E.intestinalis*) and 284 (microsporidium *E.aedis*) proteins possess such a domain. In most of the cases, the Pfam domains observed in the orphans are also represented in the fraction of proteins with orthologs (see Figure 4-6), and in only very few cases a Pfam domain is detected that is not also represented in the proteins with orthologs.

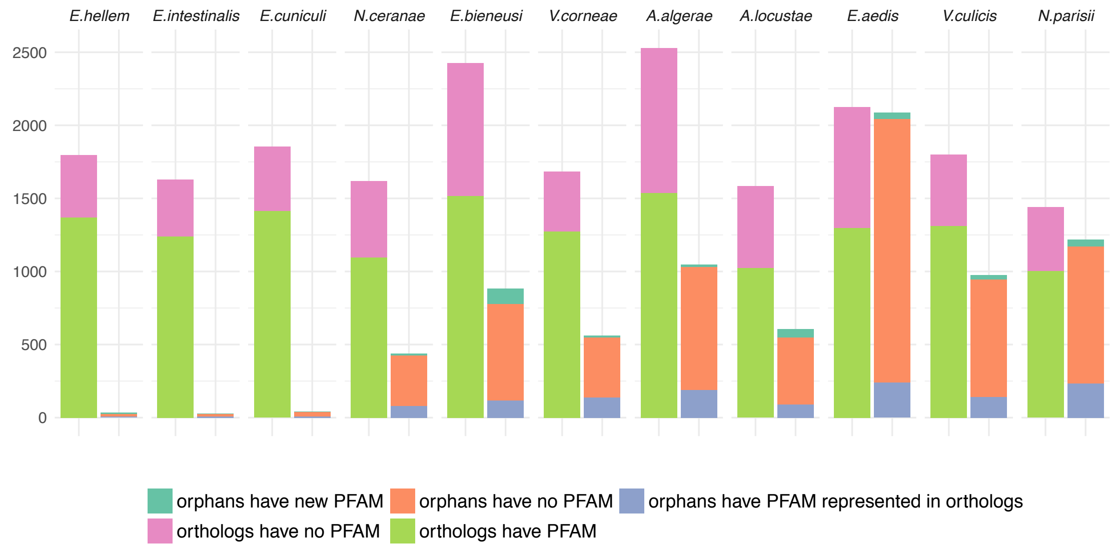

Figure 4-6: Fractions of orthologous and orphan proteins that have and do not have Pfam annotations. Each species is represented by two bars, one representing the set of proteins with orthologs in other species (left) and the other representing of orphans (right). For each set, we then identified the proteins with and without Pfam annotation. For the orphan set, we determined a third category of proteins that harbor a Pfam domain that is not observed in the proteins with orthologs (category 'new Pfam').

In summary, the microsporidian orphan proteins differ in part substantially with respect to protein length and Pfam content from their counterparts that have orthologs in other species. With respect to the evolutionary origins of these genes, there exist three alternative hypotheses. Either, these genes represent lineage specific innovations by *de novo* gene acquisitions, or they represent ancient genes that, nowadays, have been retained only on one of the analyzed microsporidian lineages, or they represent gene duplication products that, subsequent to a rapid diversification, have lost any similarity to the other duplication product.

To further address the evolutionary histories of the microsporidian proteins, we searched for orthologs in a set of 24 non-microsporidian species (Taxon set B, Table 4-2).

4.3.2 The microsporidian LCA protein set

We extended the initial orthologous groups by searching for orthologs in non-microsporidia species (Taxon set B). To this end, we used each of the orthologous groups from OrthoMCL as input for HaMStR. Out of 2904 extended groups, we could extend 1842 with sequences from taxa outside the microsporidia. Because any downstream analysis of this data set requires a robust phylogeny of our taxon collection, we first identified the subset of genes with one-to-one orthologs across the entire species set. More precisely, our two stage ortholog search revealed that each of the 11 microsporidia, and each of the 24 non-microsporidian species harbored exactly one ortholog for these genes. 80 genes met this criterion are listed in Table A-6 in Appendix. The corresponding 80 orthologous groups, each comprising 35 sequences, served then as our data basis for a phylogenomics reconstruction of our species tree.

We aligned each of the 80 orthologous groups with ClustalW v2.1, and subsequently concatenated the individual alignments into a supermatrix. This supermatrix spans 35 species and has a total length of 86,424 positions. Removing columns with more than 50% gaps shortened the super-alignment 36.616 positions. A ProtTest analysis identified the LG model of sequence evolution (Le and Gascuel 2008) modeling rate heterogeneity across sites with a Γ distribution (Parameter G), including proportions of invariable sites (Parameter I) (Steel, Huson, and Lockhart 2000) and empirical amino acid frequencies (Parameter F) as the best fitting model for the post-processed supermatrix.

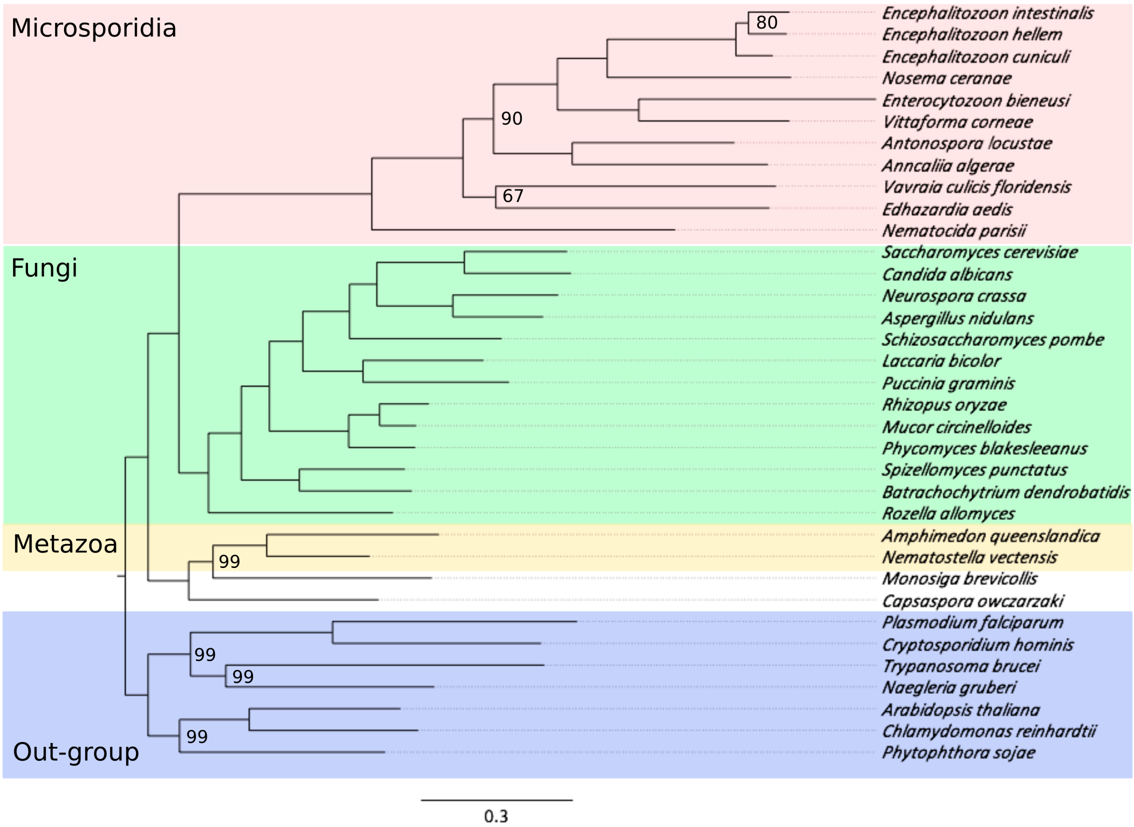

Figure 4-7: The evolutionary relationships of the species represented in taxon sets A and B. The tree is based on a supermatrix comprising 35 species and 80 genes, and spans 36,616 positions. The tree was reconstructed using the maximum likelihood criterion and the LG model of sequence evolution (see main text for further details). Branch support was assessed with 100 non-parametric bootstrap replicates, and node labels represent percent bootstrap support. The 11 microsporidia taxa are highlighted in red. As further opisthokont species, the tree represents 13 fungi (green) spanning the full phylogenetic diversity of fungi, 2 animals (yellow) alongside their close relatives, *M.brevicollis*, *C.owczarzaki*. 7 species outside the opisthokont diversity represent the outgroup of this analysis (purple). Internal node labels denote the bootstrap support and only values less than 100 are shown. The tree is rooted according to Roger and Simpson (2009).

The maximum likelihood tree reconstructed from the super-alignment and the optimal model is shown in Figure 4-7. The tree spans the full eukaryotic diversity and is overall well resolved. All but three splits achieve bootstrap support values of 99 or 100, where the splits with lower support are all within the fast evolving microsporidia. The microsporidia are placed as sister to the fungi to the exclusion of the metazoan and their close relatives, *M. brevicollis* and *C. owczarzaki*.

Based on the species tree we filtered the extended homologous groups that did not match the parsimony criteria as described in 4.2.5. Finally, we yielded 1605 final orthologous groups, which represent the set of microsporidian LCA proteins.

4.3.3 The origin of microsporidia

We attempted to resolve the position of microsporidia in the eukaryotic species tree by expending the taxon set with a more diverse set of 48 fungi from different phyla and non-fungal species as outgroup (cf. taxon set C, Table A-2 in Appendix). Here we used the microsporidian proteins in the 80 gene set (Appendix, Table A-6) as the initial core orthologous groups and searched for their orthologs in non-microsporidian species using HaMStR. The resulting supermatrix contains 72 species and has a length of 37,891 positions. The resulting maximum likelihood tree is shown in Figure 4-8. The tree was rooted using the bikont group according to (Roger and Simpson 2009). Similar to Figure 4-7, microsporidia are placed next to the fungi as the sister clade.

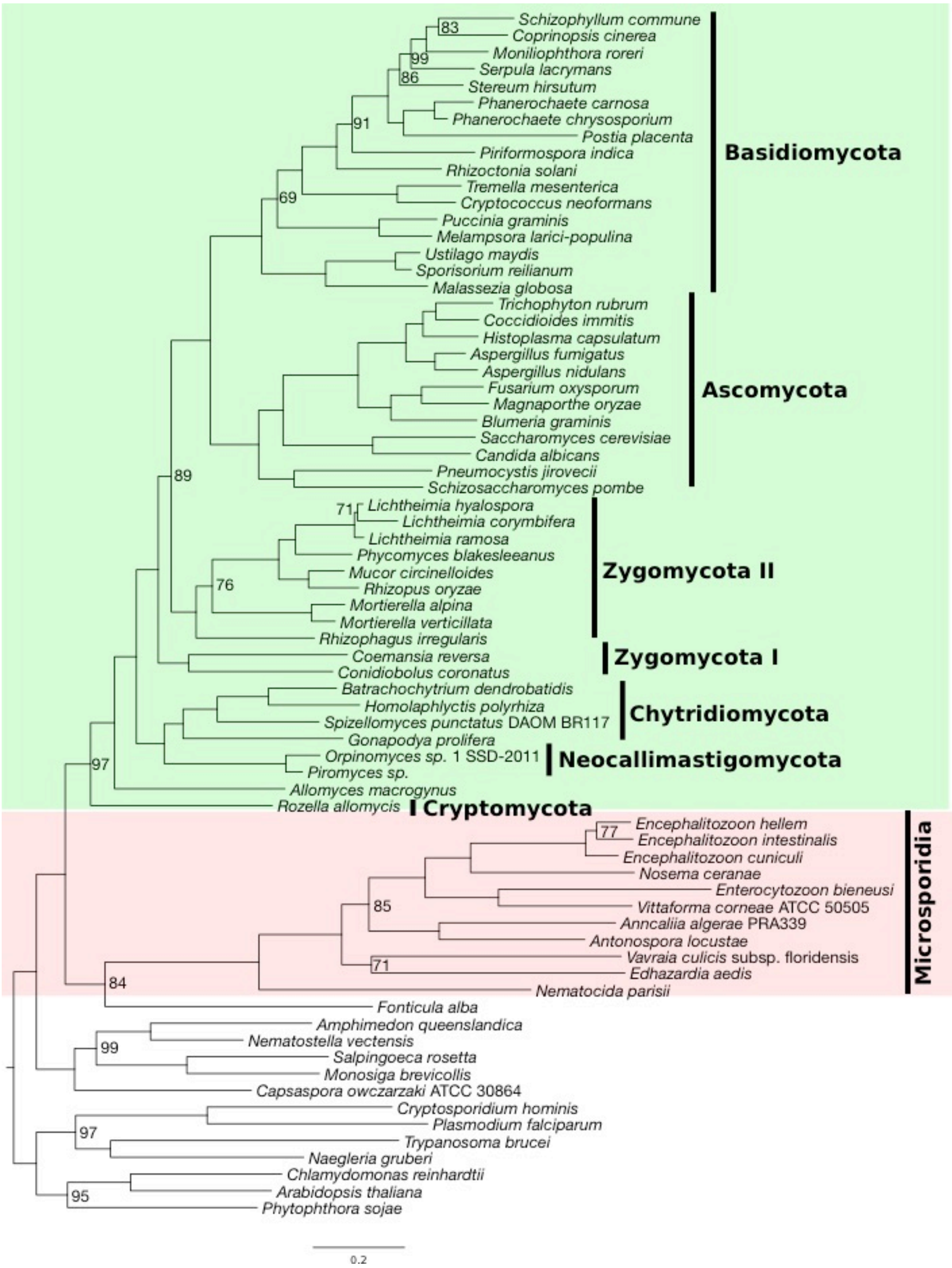

Figure 4-8: The maximum likelihood phylogeny of the fungi based on the microsporidian core gene set. Fungal taxa are highlighted in green. Microsporidian species are highlighted in red. Internal node labels denote percent bootstrap support and only values less than 100 are shown. The tree is rooted according to Roger and Simpson (2009) using the bikont group. The tree supports monophyletic opisthokonts and places the microsporidia as sister to the fungi.

We next tested whether the data contains sufficient phylogenetic information to not only support the sister group relationship of microsporidia and fungi, but to also reject the alternative hypotheses about the evolutionary origin of microsporidia that have been discussed so far. Those hypotheses placed microsporidia either on the earliest clade of all eukaryote, grouped them together with Ascomycota, Zygomycota, Cryptomycota, or placed them as the sister group of both Ascomycota and Basidiomycota. The routines for testing alternative tree topologies implemented into CONSEL revealed that the tree shown in Figure 4-8 explains the data significantly better than the alternative topologies resembling the alternative hypotheses (see Table 4-3).

Table 4-3: Result of topology tests between the alternative topologies against the reconstructed topology.

Hypothesis	au	np	bp	pp	kh	sh	wkh	wsh
Earliest eukaryote	1e-60	1e-19	0.0	7e-192	0.0	0.0	0.0	0.0
Ascomycota	2e-36	2e-14	0.0	0.0	0.0	0.0	0.0	0.0
Zygomycota	3e-31	9e-14	0.0	0.0	0.0	0.0	0.0	0.0
Cryptomycota	1e-53	8e-17	0.0	7e-89	0.0	0.0	0.0	0.0
Sister group of Ascomycota and Basidiomycota	2e-44	6e-16	0.0	0.0	0.0	0.0	0.0	0.0

au: unbiased test; np & bp: bootstrap probabilities; pp: Bayesian posterior probability; kh: Kishino-Hasegawa test; sh: Shimodaira-Hasegawa test; wkh: weighted Kishino-Hasegawa test; wsh: weighted Shimodaira-Hasegawa test

4.3.4 Phylogenetic profiles of the microsporidian LCA set

Now that we have clarified the phylogenetic position of microsporidia in the tree of life, we next investigated the evolutionary history of the genes we have assigned to the last common ancestor of the microsporidia. We traced the microsporidian LCA proteins through 491 species across the tree of life in taxon set D (Appendix, Table A-1) and analyzed their phylogenetic profiles.

To assess not only the evolutionary relationships to proteins in other species, but also assess whether or not they probably have diverged in function, we complemented the orthology assignment with FAS scores. The high FAS score frequency in Figure 4-9 revealed the high similarity in the domain architectures between the microsporidia proteins and their orthologs. Most of the protein pairs have a FAS score higher than 0.75 (mean score 0.96).

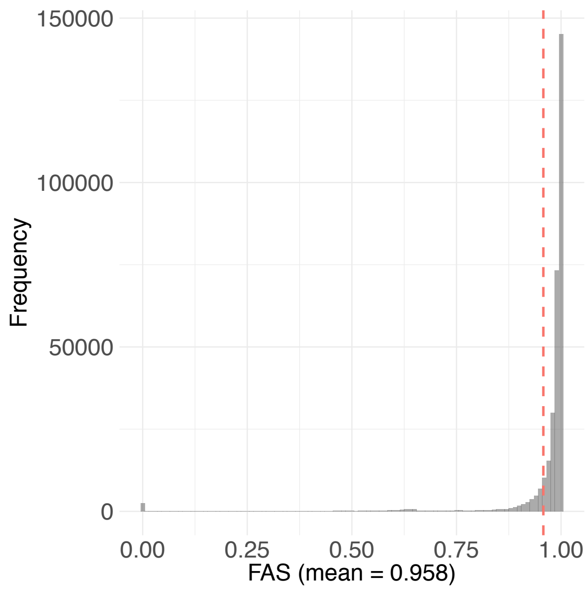

Figure 4-9: The distribution of FAS scores for all orthologs of 1605 microsporidian LCA proteins.

We clustered 1605 phylogenetic profiles of the microsporidia LCA proteins and display the whole profile plot using PhyloProfile to have an overview about their distribution. Figure 4-10 shows the complete profile across 491 taxa grouped into phylum level. It can easily be seen that a large fraction of microsporidia proteins spread through all studied taxa.

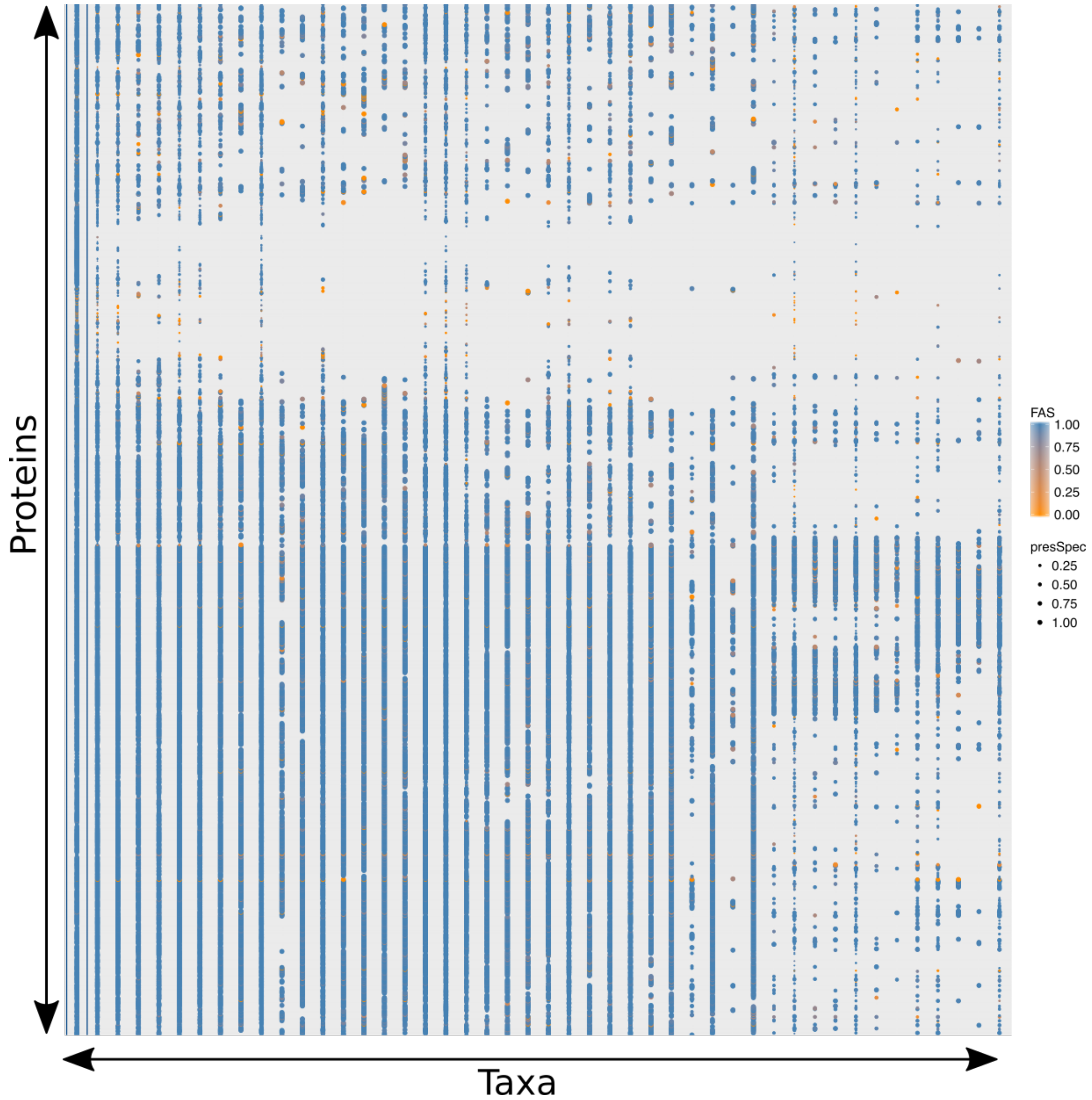

Figure 4-10: The full phylogenetic profile of 1605 microsporidian LCA protein across 491 taxa grouped in phylum level. The color of the points denotes the FAS score between microsporidia and non-microsporidia protein. The size of the points depicts the percentage of species that have orthologs in each phylum.

Based on the profiles, we then estimated the evolutionary ages for the microsporidian LCA proteins. In accord with the result in Figure 4-10, half of the proteins could be found at the root of the species tree of life and another 44% of the proteins are as old as the last eukaryotic common ancestor. Only 3% are specific to microsporidia lineage (Figure 4-11).

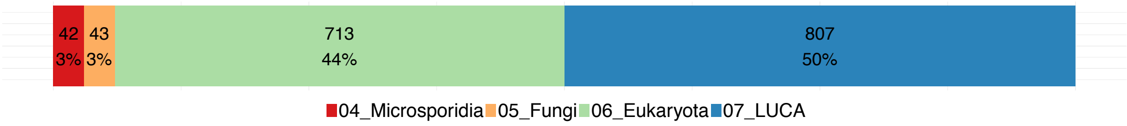

Figure 4-11: Gene age estimation of 1605 microsporidian LCA proteins. The fraction and corresponding absolute number of proteins for each estimated evolutionary age are written in each block. The colors denote the estimated ages for query proteins.

The stringency of the orthology prediction and the high FAS scores between the microsporidia proteins and their orthologs indicated that they are similar to each other not only in their sequences but also in term of functional equivalence. Because the FAS scores were already high, the result was not affected much by applying different FAS cutoffs (see Table 4-4).

Table 4-4: Estimated microsporidia specific proteins by applying different FAS cutoffs.

FAS cutoff	Microsporidia specific	LCA between microsporidia and fungi	Last eukaryotic common ancestor
0.5	3%	3%	94%
0.75	4%	3%	93%
0.9	5%	3%	92%

To investigate the functionality of those 42 microsporidia specific proteins, we used BlastKOALA (http://www.kegg.jp/blastkoala/) and HamFAS (described in Chapter 3) to annotate KEGG ids identifiers for those proteins. Only 8 of them were linked to KO groups (Table 4-5).

Table 4-5: KO annotation for 42 microsporidia specific proteins using BlastKOALA and HamFAS.

LCA protein	KO id	Description
OG_1087 (1)	K17866	Diphthamide biosynthesis protein 2
OG_1349 (2)	K18592	Gamma-glutamyltranspeptidase
OG_1378 (3)	K09485	Heat shock protein 110kDa
OG_1378 (1)	K09489	Heat shock 70kDa protein 4
OG_1515 (1)	K08803	Death-associated protein kinase
OG_1710 (1)	K14848	Ribosome assembly protein RRB1

OG_1710 (2)	K04802	Proliferating cell nuclear antigen
OG_2013 (3)	K02155	V-type H+-transporting ATPase 16kDa proteolipid subunit
OG_2250 (3)	K02896	Large subunit ribosomal protein L24e
OG_2280 (3)	K02180	Cell cycle arrest protein BUB3

(1) *protein was annotated only by BlastKOALA*

(2) *protein was annotated only by HamFAS*

(3) *protein was annotated by both BlastKOALA & HamFAS*

Beside KO annotation, we classified the microsporidia specific proteins based on Gene Ontology terms (Ashburner et al. 2000) using Blast2GO v5.0.13 (Götz et al. 2008). Additionally, 12 other microsporidian LCA genes were annotated by GO terms (Figure 4-12, or more detail in Appendix, Table A-4).

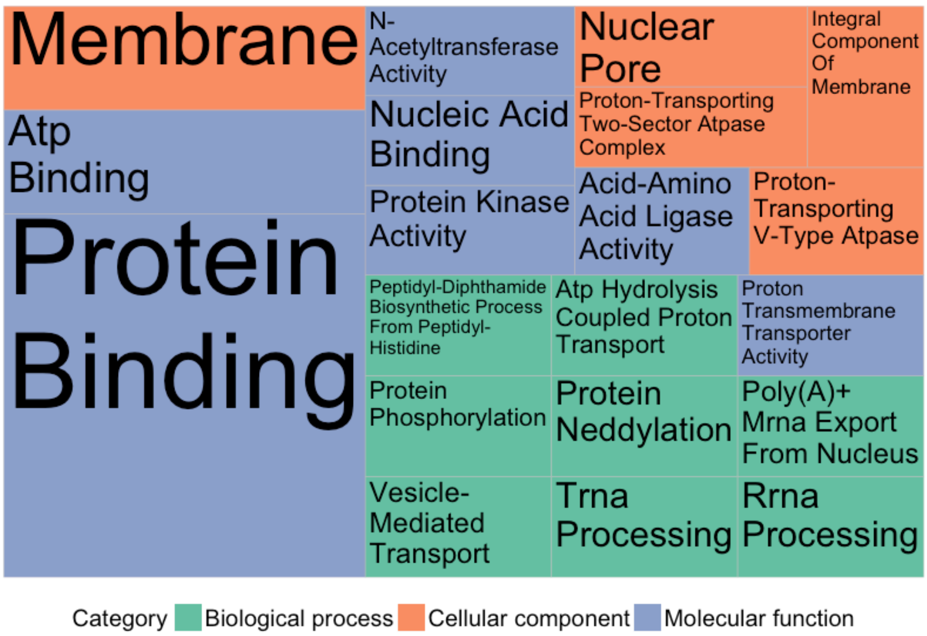

Figure 4-12: GO annotation for microsporidia specific proteins.

4.3.5 The metabolic characteristics of the microsporidian LCA

In this section we analyze the characteristics from the microsporidian last common ancestor, as they can be reconstructed from then LCA gene set, and compare it to the contemporary species.

KO annotation of the microsporidian LCA proteins

Using HamFAS we annotated 1344 distinct KO identifiers to 1048 of the 1605 microsporidian LCA proteins.

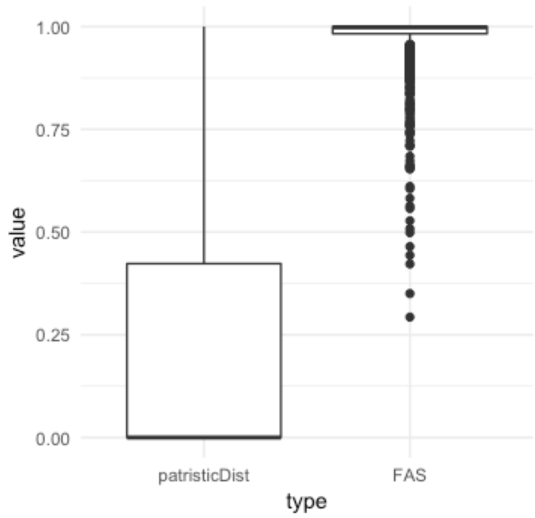

Figure 4-13: Distribution of FAS scores and patristic distances of KO-annotated microsporidian LCA proteins.

The distribution shown in Figure 4-13 revealed a trend of high FAS scores (mean FAS score is 0.97) and low patristic distances (mean and median are 0.22 and 0.00 respectively) for a large fraction of annotated KO ids.

We furthermore analyzed the metabolism of the microsporidian LCA by mapping its proteins into different reference KEGG pathways.

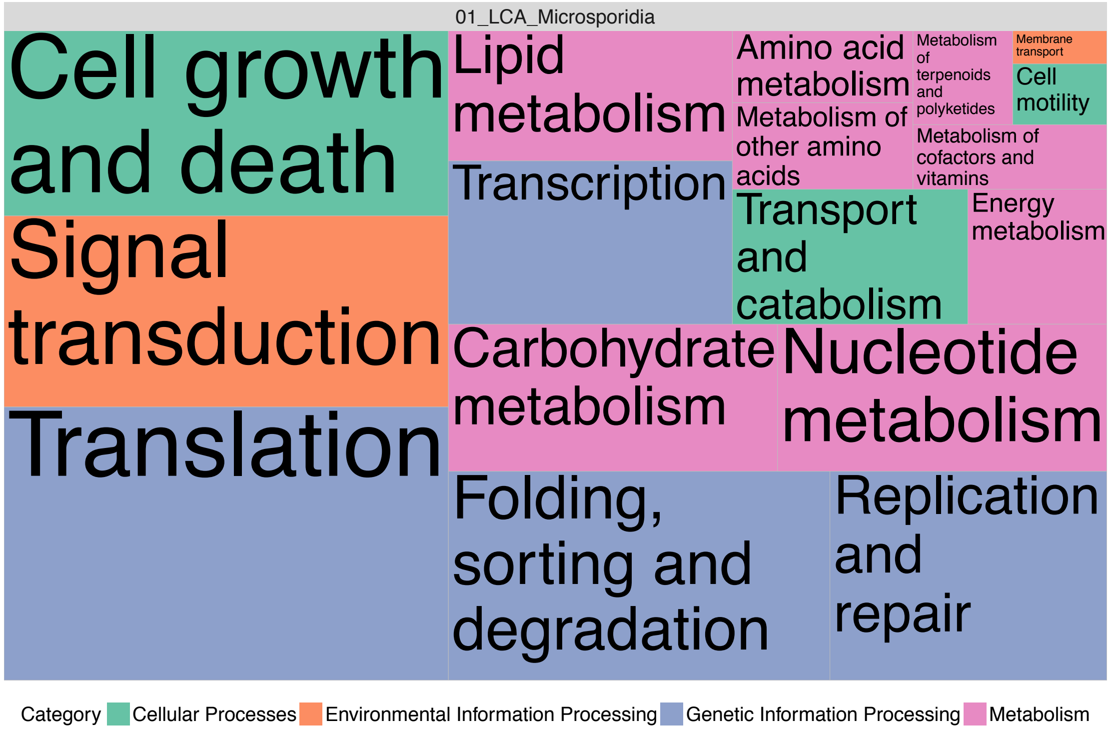

Figure 4-14: The distribution of microsporidia LCA proteins in different pathway categories - Cellular processes (green), environmental information processing (orange), genetic information processing (purple) and metabolism (pink).

The relative fractions of microsporidian LCA proteins distributing in different pathway categories is displayed in Figure 4-14. The largest fraction is of the genetic information processing proteins, which comprises 42% of the mapped proteins. The other 30% belong to metabolic pathways, which is higher than we observe in extant microsporidia species (25% in average) but much lower than the free-living *S.cerevisiae* with 38% (see Appendix, Figure A-7).

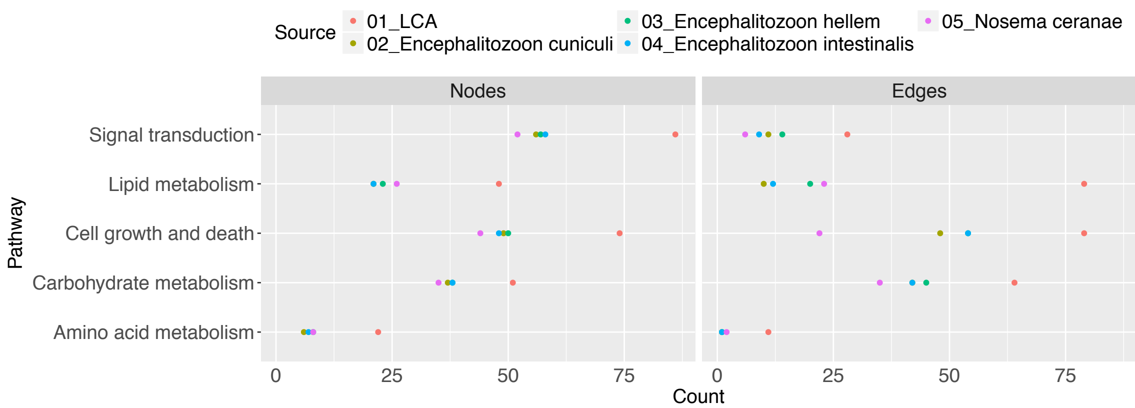

Figure 4-15: Number of nodes (left) and edges (right) of the enriched pathways for microsporidian LCA, *E.cuniculi*, *E.hellem*, *E.intestinalis* and *N.ceranae*.

In particular, the proteins assigned to the microsporidian LCA are preferentially involved in carbohydrate, amino acid and lipid metabolism, cell growth and death, signal transduction, folding, sorting and degradation of proteins. The enrichment of the LCA's proteins in those pathways in comparison to the one of four contemporary microsporidia is shown by the higher number of nodes and edges in the connectivity network (Figure 4-15). The average node degree, average path length and the diameter (the longest shortest path) in Figure 4-16 reveal a highly connecting grade of the microsporidian LCA proteins in comparison to other contemporary species.

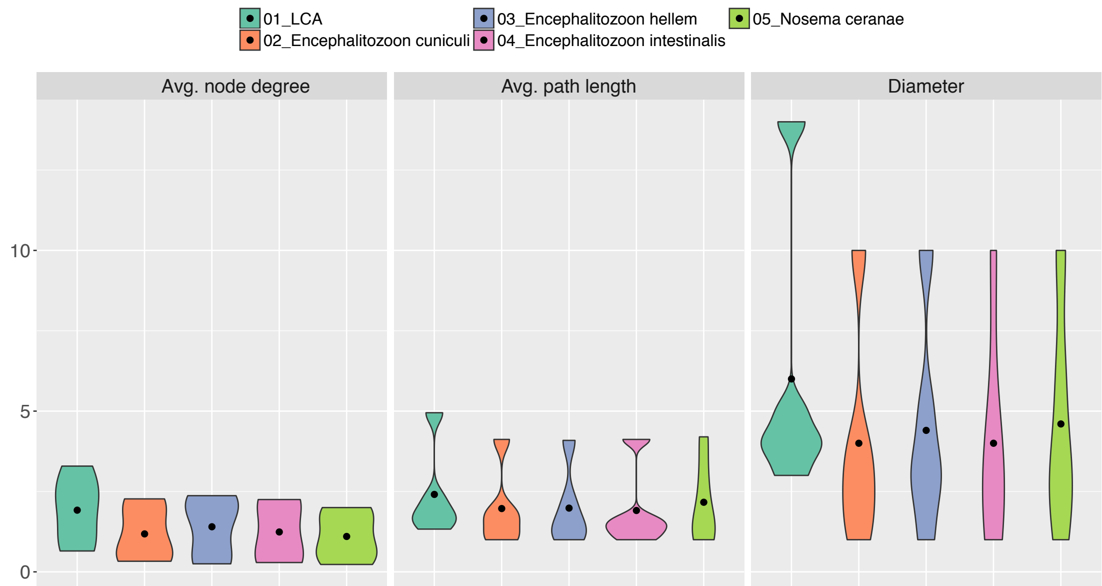

Figure 4-16: Density of average node degree, average path length and diameter (maximal path length) of microsporidian LCA, *E.cuniculi*, *E.hellem*, *E.intestinali* and *N.ceranae* in the enriched pathways (amino acid, carbohydrate, lipid metabolism; cell growth and death; signal transduction; folding, sorting and degradation). The higher mean scores of the LCA indicate the higher connecting grade of the LCA's proteins in the pathways in comparison to other species.

The evidence for the mitochondrial ancestry of the microsporidia

According to several studies e.g. (Fast and Keeling 2001; Keeling and Fast 2002; Agnew et al. 2003), microsporidia lack mitochondria. The presence of genes coding for the mitochondrial heat-shock protein 70 (hsp70) in some extant microsporidia species, suggests that microsporidia ancestor had mitochondria though (Germot, Philippe, and Le Guyader 1997; Hirt et al. 1997). Those studies also hypothesized that microsporidia replaced the pyruvate dehydrogenase complex (PDH) by pyruvate ferredoxin oxidoreductase (PFOR) in order to convert pyruvate into acetyl-CoA and produce NADH. However, we could not annotate any KO id of the PFOR subunits (α, β, γ, δ) for the microsporidian LCA proteins. Instead, two out of three components of PDH were found, namely the pdhA and pdhB of E1 component, as well as E3 (DLD) component (cf. Appendix, Table A-7). The E2 (DLAT, K00627) component was not found. Note that E1 is also be found in *N.locustae* (Fast and Keeling 2001) and the genus *Encephalitozoon* (Katinka et al. 2001). Figure 4-17 shows the mapped microsporidian LCA proteins into the pyruvate decarboxylation process.

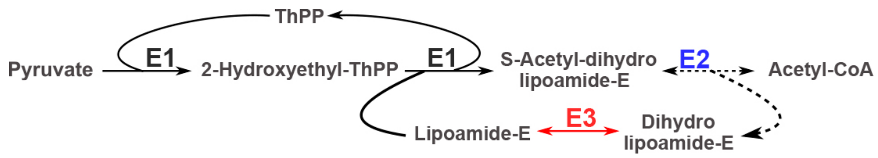

Figure 4-17: The process converts Pyruvate into Acetyl-CoA with the help of pyruvate dehydrogenase complex (PDC). Beside the E1 component, which was found in the extant species, the microsporidia LCA has in additional the E3 component (red). E2 (blue) is the missing component in both LCA and contemporary microsporidia.

The lack of TCA cycle and its replacement

Microsporidia lack mitochondrial pathways and reactions, such as electron transport chain, oxidative phosphorylation pathway and TCA cycle (Keeling and Fast 2002; Keeling 2009; Wiredu Boakye et al. 2017). All the required enzymes for TCA are missing in both the LCA and all contemporary microsporidia. Likewise, they retain only 10/13 subunits of the vacuolar H^+ ATPase in the oxidative phosphorylation.

Due to the lack of the main ATP supplier from the mitochondrion, the synthesis of ATPs therefore depends on other pathways like glycolysis or through ATP transport system (Keeling and Fast 2002; Keeling and Corradi 2011; Heinz et al. 2012).

Microsporidia mostly take up ATP from the host species using their ATP-binding cassette (ABC) transporters (Méténier and Vivarès 2001; Keeling 2009; Heinz et al. 2012). Besides, (Heinz et al. 2012) also reported putative major facilitator superfamily (MFS) transporters in the microsporidia *T.hominis*. We searched for those transport proteins in the microsporidian LCA and found two MFS transporter and 6 ATP-binding cassette (ABC) transporters (see Figure 4-7).

Table 4-6: Microsporidian LCA MFS and ABC transporters.

LCA protein	KO id	Description
OG_3349	K08139	MFS transporter, SP family, sugar:H+ symporter
OG_1075	K08146	MFS transporter, SP family, solute carrier family 2 (facilitated glucose transporter), member 9
OG_1019	K06174	ATP-binding cassette, sub-family E, member 1
OG_1050	K06185	ATP-binding cassette, subfamily F, member 2
OG_1034	K06158	ATP-binding cassette, subfamily F, member 3
OG_1082	K05681	ATP-binding cassette, subfamily G (WHITE), member 2
OG_1098	K05662	ATP-binding cassette, subfamily B (MDR/TAP), member 7

The microsporidian LCA's carbohydrate metabolism

Besides the presence of enzymes responsible for glycolysis, the annotation of the microsporidia LCA proteins also suggested that it has also the pentose phosphate pathway, another core carbon metabolism that has been found in the contemporary microsporidia (Keeling and Fast 2002; Keeling and Corradi 2011; Heinz et al. 2012).

The primary carbohydrate storage trehalose is thought to be very essential for the survival and germination of microsporidian spore (Vandermeer and Gochnauer 1971; Dolgikh, Sokolova, and Issi 1997; Agnew et al. 2003; Heinz et al. 2012). Enzymes for trehalose synthesis and degradation in extant microsporidia (Vandermeer and Gochnauer 1971; Méténier and Vivarès 2001; Keeling and Corradi 2011; Heinz et al. 2012) have also been found in the LCA including the trehalose 6-phosphate synthase and alpha-trehalase (cf. Appendix, Table A-7).

Figure 4-18 demonstrates the scheme of possible carbohydrate metabolism of the microsporidian LCA.

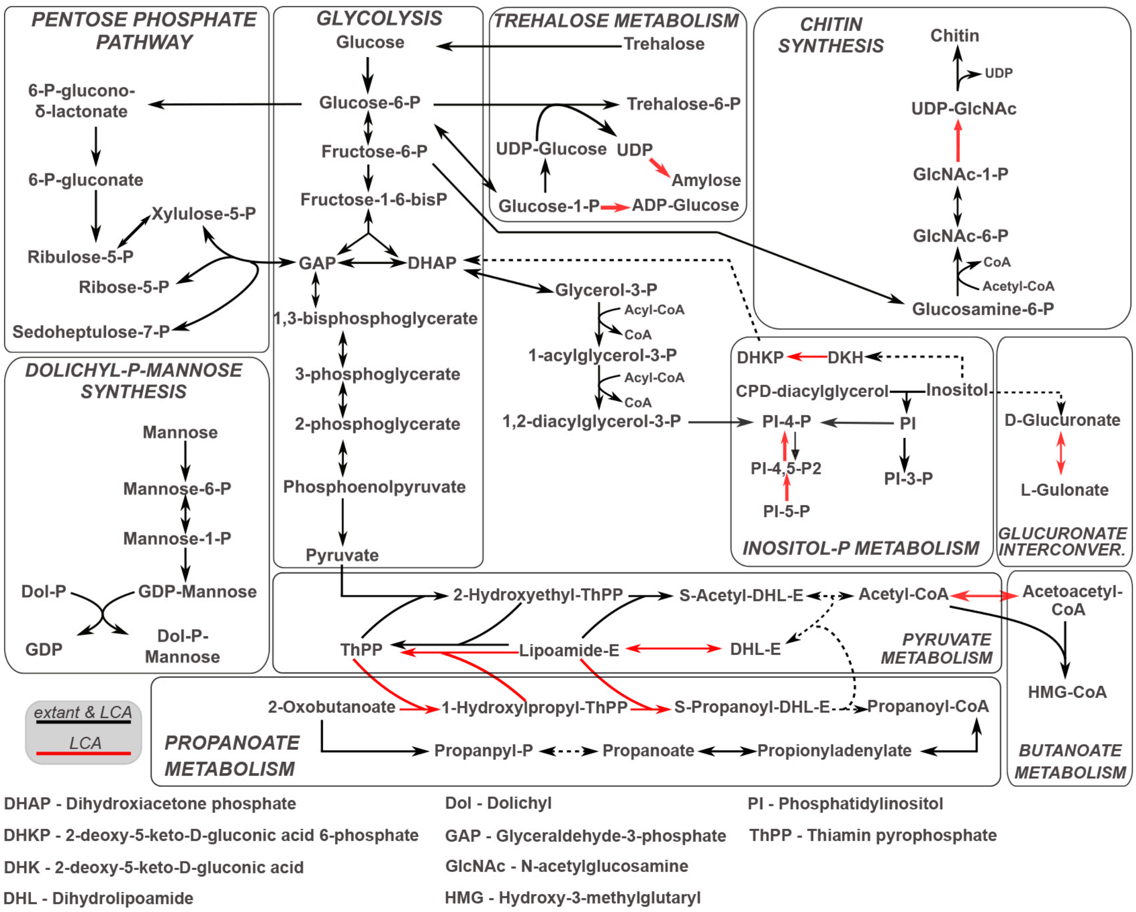

Figure 4-18: Scheme of the carbohydrate metabolism in microsporidia. The solid black arrows represent reactions that present in both microsporidian LCA and extant species. Red arrows are reactions, whose enzymes are found only in the LCA. The dashed black arrows indicate missing reactions.

The inability for nucleotide synthesis in microsporidia

As obligate intracellular parasites, microsporidia have the option to uptake nucleotide from the host than produce it by themself (Heinz et al. 2012; Dean, Hirt, and Embley 2016). Just as the extant species, the microsporidian LCA lacks ribose-phosphate pyrophosphokinase (K00938, EC 2.7.6.1), IMP cyclohydrolase (K11176, EC 3.5.4.10) and UMP synthetase (K13421, EC 2.4.2.10 & 4.1.1.23), which are key enzymes for the de-novo purine and pyrimidine synthesis. Those enzymes involve in converting ribose 5-phosphate into phosphoribosyl pyrophosphate (PRPP), and synthesizing inosine monophosphate IMP and UMP from PRPP. Figure 4-19 describes the nucleotide metabolism of the microsporidian LCA and the contemporary species.

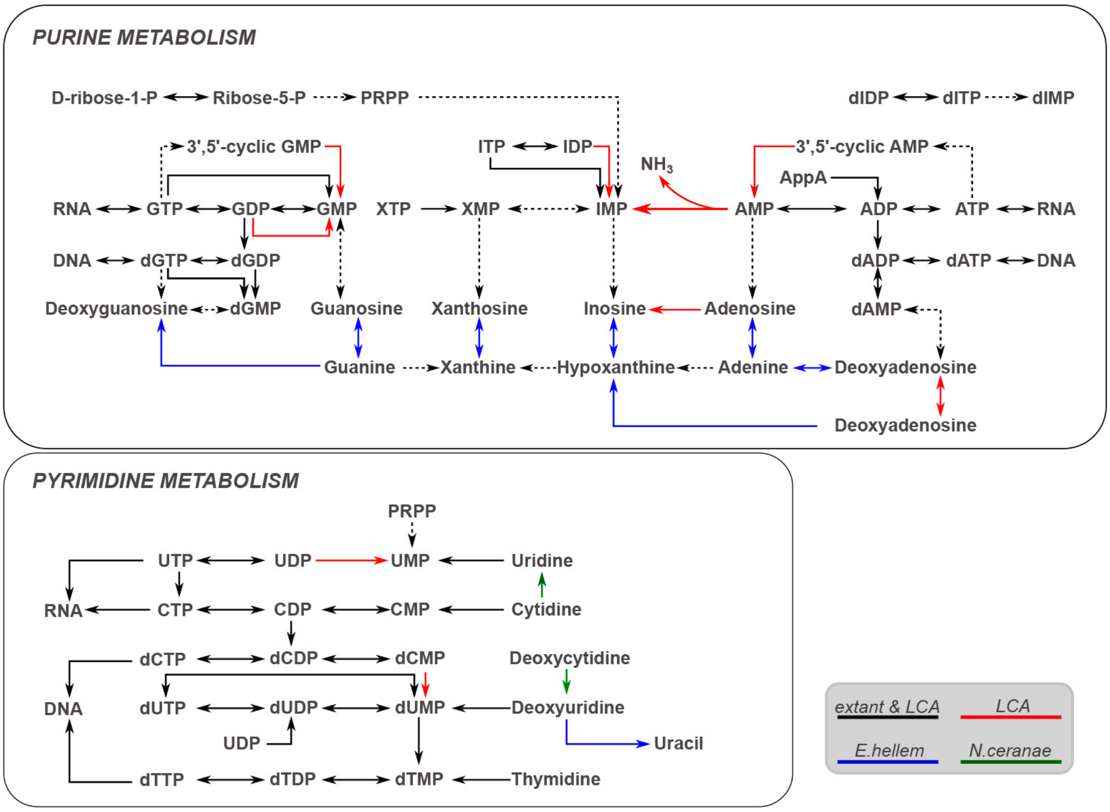

Figure 4-19: Scheme of nucleotide metabolism in microsporidia. The solid black arrows represent reactions that present in both microsporidian LCA and extant species. Red, blue and green arrows are reactions, whose enzymes are found only in the LCA, *E.hellem* and *N.ceranae* representatively. The dashed black arrows indicate missing reactions.

For that reason, microsporidia need to import nucleotides from their hosts using nucleotide transport (NTT) proteins. The KO id K03301 of four NTT (NTT1, NTT2, NTT3, NTT4) proteins (Heinz et al. 2014; Dean, Hirt, and Embley 2016) have also been found in the microsporidian LCA proteins (cf. Appendix, Table A-7).

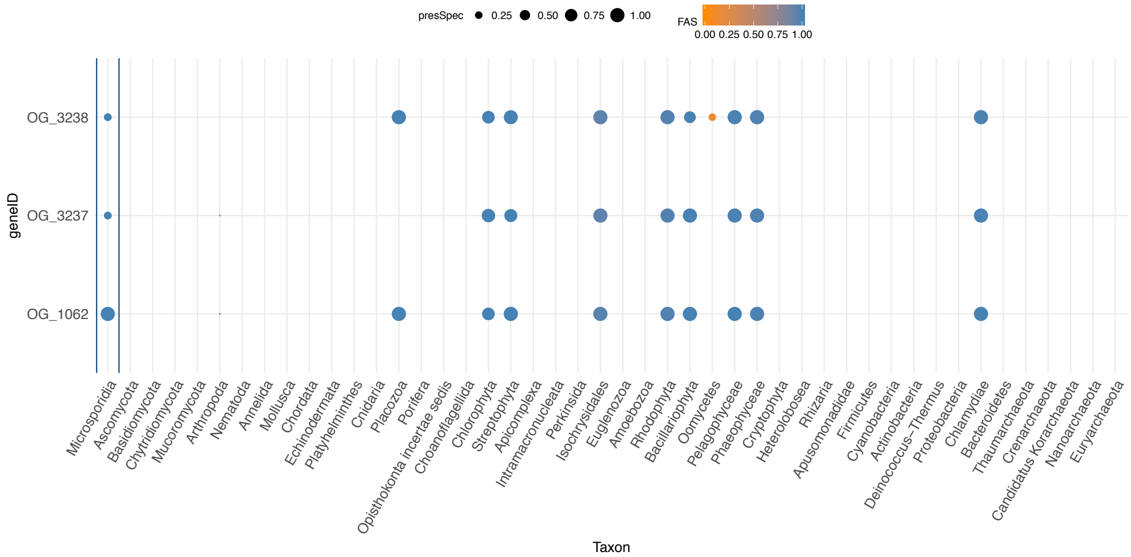

Figure 4-20: Phylogenetic profile of 3 microsporidian LCA NTT proteins.

Figure 4-20 shows the phylogenetic profile of 3 microsporidian LCA NTT proteins. All three proteins have orthologs in the bacterial phylum Chlamydiae and some other eukaryotic phyla with very high FAS scores. The domain annotation of a microsporidian protein in comparison with its bacterial ortholog is shown in Figure 4-21. They both contain 11-12 transmembrane domains, as commonly observed in bacterial NTT proteins (Winkler and Neuhaus 1999; Tsaousis et al. 2008).

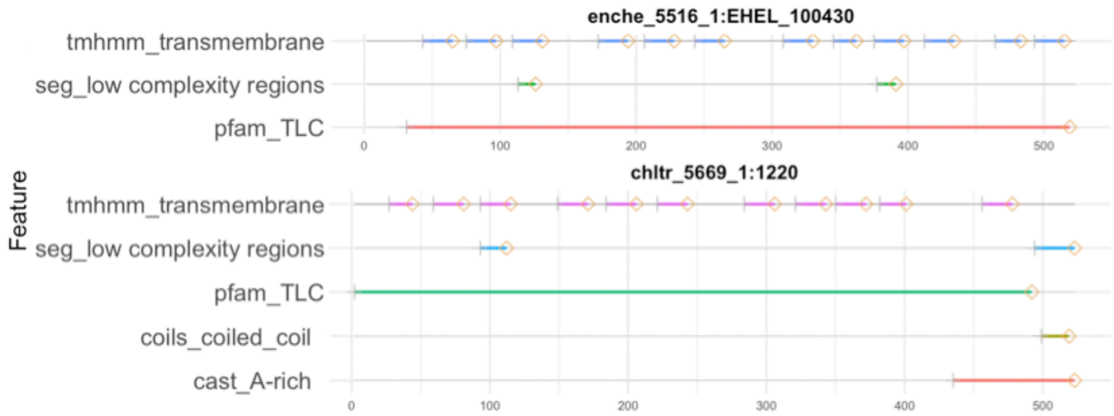

Figure 4-21: Domain architecture of *E.hellem* protein (enche_5516_1:EHEL_100430) and its ortholog (chltr_5669_1:1220) of the bacteria *Chlamydia trachomatis*.

Ancient reactions in the microsporidian LCA's metabolism

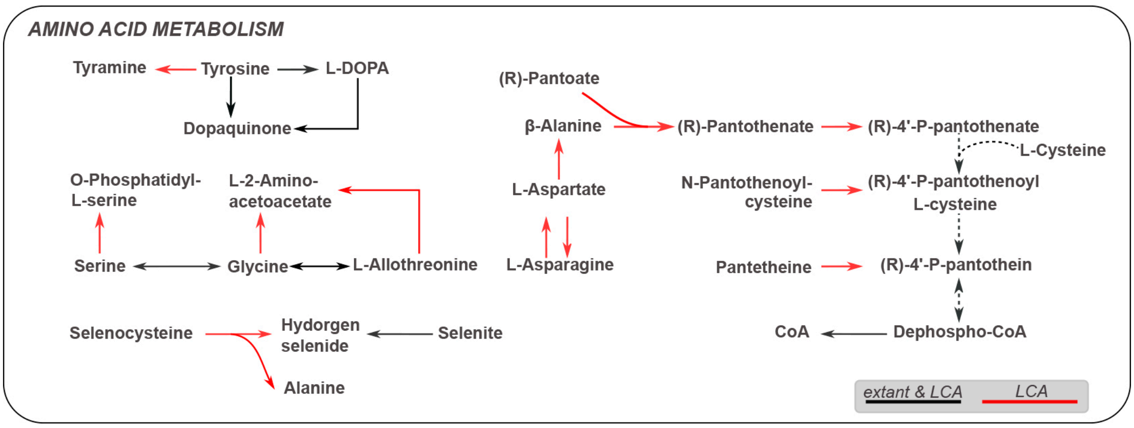

Figure 4-22: Reconstruction of the amino acid metabolisms in the microsporidian LCA. Red arrows indicate reactions that could be found only in the LCA, while solid black arrows are the one present in both LCA and extant microsporidia. Dashed black arrows are missing reactions.

The schematic metabolisms for carbohydrates in Figure 4-18 as well as for purines and pyrimidines in Figure 4-19 unveiled some ancient reactions in the microsporidia LCA in comparison to the extant species, whose metabolic pathways are available in KEGG database. The ancient reactions were also identified in some amino acid metabolism (Figure 4-22), glycerophospholipid metabolism (Figure 4-23), and other cellular process and genetic information processing pathways (Appendix, Figure A-8, Figure A-9, Figure A-10). These ancient reactions can fill the gaps in those pathways and could, therefore, suggest a more effective utilization of those pathways of the microsporidian

LCA than the 4 contemporary species in this study. The corresponding proteins of those ancient reactions serve as potential candidates for further experimental analyses of microsporidian metabolism.

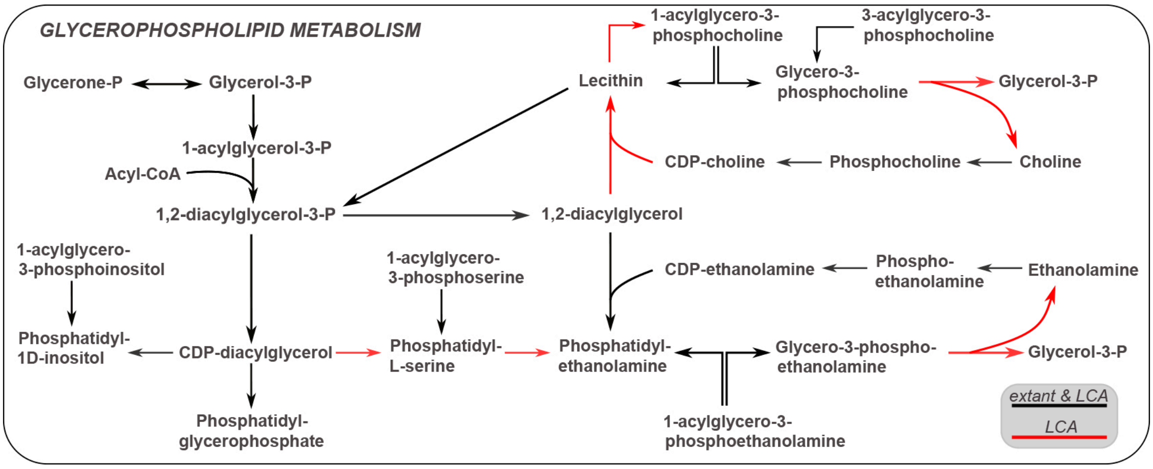

Figure 4-23: Scheme of glycerophospholipid metabolism in the microsporidia LCA. Red arrows indicate reactions that could be found only in the LCA, while solid black arrows are the one present in both LCA and extant microsporidia.

4.4 Discussion

4.4.1 The evolutionary history of microsporidian proteins

It has been shown that, even with the pronounced genome reduction, microsporidia still have a fraction of species-specific genes (Cuomo et al. 2012; Peyretaillade et al. 2012). We found between 2% of the predicted proteins in the genus *Encephalitozoon* and up to 49% of proteins in *E.eadis* were microsporidian orphan proteins. Different portions of orphan genes suggested an independent gain and loss in genomes of microsporidia species. Moreover, the number of genes that are shared between microsporidia are not correlated with the genome size. For instance, *E.aedis*, whose genome size is 25 folds larger and the number of predicted genes is two times as large as the one of *E.cuniculi*, has only 1.14 times more shared proteins than *E.cuniculi*. Thus, it declines the possibility of whole genome duplication hypothesis in some microsporidia lineages.

These orphan genes are thought to be newly invented genes that are related to the parasitism of microsporidia and could help to identify the host range for each species (Vivarès and Méténier 2001; Hirt and Horner 2004). We performed a Pfam annotation for those proteins. However, consistent with Cuomo et al. (2012), a large fraction of those orphan genes, 75% in average, were not assigned any Pfam domains. In most cases where the annotation was success, the Pfam domains observed in the orphans are also represented in the fraction of proteins with orthologs. This indicated that, those sequences have been evolved so quickly that OrthoMCL was not sensitive enough to find their ortholog in other species. On the other hand, the length comparison between orphan and orthologous proteins however suggested that the orphan sequences could also be the artifact of the gene prediction processes. As wrongly predicted genes has also been observed by Peyretaillade et al. (2012), further analyses are needed to distinct between those genes and the short orphan genes in microsporidia species.

Due to the compact genomes of the extant microsporidia taxa, most of the proteins in the microsporidian LCA were evolutionary old (Nakjang et al. 2013). Not surprisingly, the phylogenetic analysis in 4.3.4 showed that, 94% of the microsporidian LCA proteins are evolutionary old and can be traced back to the last eukaryotic common ancestor. Another 3% of the proteins share the common ancestor with fungal clade. It remains only 3% LCA proteins that are specific to the microsporidia lineage. Although the KO and GO annotations could not provide much knowledge about the functions of those microsporidia specific proteins, they were supposed to play an important role in the interactions between microsporidia and the host species for adapting with their parasitic lifestyle (Nakjang et al. 2013). We filtered the ortholog inference using different cutoffs of protein feature architecture similarity (FAS) scores. However, this filtering step did not change the result much, since the mean FAS score of the original set of orthologs was already high (0.958).

This outcome consists with the dynamic evolution of microsporidian genomes, in which they underwent not only the reduction but also the expansion process to adapt to their parasitic lifestyle (Nakjang et al. 2013).

4.4.2 The microsporidian origin

Due to the compact genomes and lack of several typical eukaryotic cellular components, microsporidia were classified as one of the earliest eukaryote taxa placing in the phylum Archezoa based on some electron microscopy as well as phylogenetic studies (Kudo and Daniels 1963; Vossbrinck et al. 1987). Recently, more and more evidences supported the fungal related origin of microsporidia (Hirt et al. 1999; Fast and Keeling 2001; Capella-Gutiérrez, Marcet-Houben, and Gabaldón 2012; James et al. 2013). Nonetheless, the exact relationship between microsporidia and fungi is still debated (Stentiford et al. 2016). Our phylogenetic tree in 4.3.3 - reconstructed from the microsporidia core gene set - strongly support the hypothesis that the microsporidia are the sister clade of the fungi. We increased the taxon sampling to avoid typical artifacts of phylogenetic inference (Zwickl and Hillis 2002; Bergsten 2005), especially to reduce the effect of long branch attraction on the reconstructed phylogeny of microsporidia (Keeling and Fast 2002; James et al. 2013). The taxon sampling comprises of eleven microsporidia from the most compact *Encephalitozoon intestinalis* with 1657 proteins to *Edhazardia aedis* with more than 4000 proteins, a diverse set of fungi from different phyla together with other six opisthokonts and a group of seven bikonts. According to the statistical tests from CONSEL in 4.3.3, our data supports the tree topology where microsporidia are placed as the sister group or fungi significantly better than other hypotheses, such as the earliest clade of eukaryote (Cavalier-Smith 1983), sister group of Ascomycota and Basidiomycota (Gill and Fast 2006), or microsporidia are close relative to Cryptomycota (James et al. 2006), Ascomycota or Zygomycota (Keeling, Luker, and Palmer 2000; Keeling 2003; Lee et al. 2008).

The sister clade of fungi origin has been also reported in the study of Capella-Gutiérrez, Marcet-Houben, and Gabaldón (2012). However, these authors, proposed this hypothesis based on a phylogenetic tree that contained only three clades, namely the microsporidia, fungi and a group of other opisthokonts. That tree, therefore, was not sufficient for such conclusion, as the position of microsporidia in the species tree is still being discussed. Our inferred tree, in contrast, was grouped using bikonta as outgroup. Thus, it can clarify the microsporidian - fungal relationship better.

4.4.3 The metabolism of the microsporidian LCA

The seed and reference proteins are highly similar to each other in term of domain architectures and a large fraction of the annotations come from the less divergent ortholog sequences to the seed proteins. This confided the transferred KO annotations of the microsporidian LCA proteins.

In general, microsporidian LCA has more proteins mapped into KEGG pathways in comparison to extent microsporidia species. However, it is still much less compared to *S.cerevisiae*, a representative of free-living organisms (see Appendix, Figure A-7). This is congruent with the reduction hypothesis of microsporidia genomes (Luallen et al. 2016). Beside the biological reason, technically this was an arbitrary comparison, since the number of yeast proteins in this analysis is much higher than the one from microsporidia (3534 yeast proteins versus 1000 proteins in average for each microsporidia species).

The origin of mitochondria in the microsporidian LCA was repeatedly discussed (Germot, Philippe, and Le Guyader 1997; Hirt et al. 1997). We could confirm their ancestral presence through the annotated LCA proteins with the presence of E1, E3 components and the hsp70 proteins. However, the role of those proteins is still unclear (Fast and Keeling 2001).

Our study agreed with the assumption that microsporidia are unable to *de novo* synthesize both purines and pyrimidines and they replace that inability by the nucleotide transport (NTT) proteins (Heinz et al. 2014; Dean, Hirt, and Embley 2016). The phylogenetic profile of three microsporidian LCA NTT proteins was consistent with the study of (Nakjang et al. 2013), where we found orthologs for those microsporidia NTT proteins also in the same phyla that were discussed in that analysis, namely Chlamydiae, Streptophyta, Chlorophyta and Bacillariophyta. The NTT orthologs have no signal peptide and contain 10-12 transmembrane domains. Based on studies of Tsaousis et al. (2008); Heinz et al. (2014); Dean, Hirt, and Embley (2016), those NTT proteins are the result of horizontal transfer event from bacteria. The presence of NTT proteins to replace the inability of de novo nucleotide synthesis is an important characteristic of obligate intracellular parasitic lifestyle (Nakjang et al. 2013; Major, Embley, and Williams 2017).

The analysis of microsporidian LCA metabolic pathways acquired the consistent results with other studies. Microsporidian LCA, as well as the contemporary species, obligatory depends on the host species for their survival due to their reduced metabolism (Agnew et al. 2003; Luallen et al. 2016). The presence of transport proteins supplements the lack of some main pathways for producing energy and other important compounds (Méténier and Vivarès 2001; Heinz et al. 2012). Trehalose again has been shown to be the main carbohydrate storage for microsporidia (Vandermeer and Gochnauer 1971; Méténier and Vivarès 2001; Keeling and Corradi 2011; Heinz et al. 2012), since the enzymes for *de novo* trehalose synthesis and degradation were also found in the LCA. However, the reason for the existence of mitochondria is still unclear (Keeling 2009), since the pathways that take place in mitochondria are already missing in the microsporidian LCA.

Ancient reactions, which have lost in the extant microsporidian lineages, have been observed in the mapped pathways of microsporidian LCA. They imply a relevant complementation for those related pathways. However, since some key enzymes were missing, which hinder the in vivo synthesis of critical metabolites such as purines and pyrimidines, we suppose that the parasitic lifestyle already occurred in the microsporidian LCA.

5 Conclusion & Outlook

Microsporidia serve as a good candidate for studying the compactness of eukaryotic parasites in both genomic and metabolic aspects (Reinke and Troemel 2015). Our study on the pan-gene sets of eleven contemporary microsporidia explained the dynamic evolutionary history of their genomes (Agnew et al. 2003; Nakjang et al. 2013). In some extreme compact species such as the members of *Encephalitozoon* genus, they still have about 2% of species specific genes. This number of orphan genes can be increased up to 49% in the larger species, the *E.eadis*. The orphan genes appear to be significantly smaller than the one of genes that have orthologs and our Pfam analysis supports that these might be newly invented genes in microsporidian lineage. During the evolutionary development of each individual microsporidia species, they created new genes to adapt to their host specific parasitic lifestyle (Vivarès and Méténier 2001; Hirt and Horner 2004). In contrast to the new orphan genes, the other genes that are evolutionary old that can be traced in almost all species in the tree of life (Nakjang et al. 2013). This assumption was clearly been observed through the phylogenetic profiles of the microsporidian last common ancestor over 480 species in three domains of life. Up to 94% of the microsporidian LCA proteins are as old as the last eukaryotic common ancestor. It lefts only 3% of the proteins that have fungal orthologs and 3% are microsporidian specific proteins. This result also confirmed that, most of the genes retained in the microsporidian LCA are essential for the survival and development of microsporidia, even though they have effectively reduced the genomes while becoming obligate intracellular parasites (Agnew et al. 2003; Nakjang et al. 2013). The *de novo* invented genes of microsporidia, however still a challenge for a comprehensive understanding about microsporidian diversity (Cuomo et al.

2012), due to the fact that they are poorly functionally described. This inhibits a wider use of microsporidia as model organisms.

The metabolic pathway reconstruction of the microsporidian LCA supports the ancestral state of the obligate endoparasitic lifestyle. We found ancient reactions that could complement some gapped pathways, which are missing in the four contemporary microsporidia available in KEGG. The corresponding proteins for those reactions are hypothesized to be secondary lost during the course of evolution. Nevertheless, the lack of main pathways for energy metabolism such as the citric acid cycle or primary enzymes for biosynthesis the initial substrates for the *de novo* nucleotides metabolism proposed the dependency of the microsporidia LCA on the outside resources. A number of transporter proteins for uptaking the main energy molecule ATPs (Alberts et al. 2002a) or one of the fundamental elements for all living organisms - the nucleotides (Liu 2007) found in the microsporidia LCA furthermore supported this hypothesis.

To assess the origin of microsporidia, we used an extensive taxon sampling. The phylogenetic tree was reconstructed using a set of 80 microsporidian core genes (cf. Table A-6). We rooted the tree using a group of bikont taxa, as suggested in the study of Roger and Simpson (2009). With a diverse set of 48 fungi, we were able to test all the debated hypotheses about the microsporidia - fungal relationship (Keeling, Luker, and Palmer 2000; Keeling 2003; Gill and Fast 2006; James et al. 2006; Lee et al. 2008; Capella-Gutiérrez, Marcet-Houben, and Gabaldón 2012). Our reconstructed tree strongly supported the hypothesis that microsporidia is place as the earliest clade of fungi. The statistical tests from CONSEL rejected all other positions of microsporidia in the species tree that are proposed by other studies with the significant P-values $\ll 0.05$.

This work demonstrates a phylogenomics approach to study the evolutionary history of the microsporidia proteins, reconstruct their last common ancestor gene set and further investigate their metabolic network to gain insights into their obligate intracellular parasitic lifestyle. We showed the practicality of our developed tool PhyloProfile in exploring the complex phylogenetic profiles. Additionally, the novel annotation transfer approach HamFAS also showed to be potential for *in silico* functionally describing uncharacterized proteins. With the ability of searching orthologs in remote species using hidden Markov model profiles and the awareness of domain architecture similarity for increasing the confidence of functional equivalence, the inferred orthologs and the seed species are likely to have similar functions. Furthermore, the identified 80 microsporidian core gene set provides a promising gene collection for reconstructing deep branches in the eukaryotic phylogeny.

References

Abascal, Federico, Rafael Zardoya, and David Posada. 2005. "ProtTest: Selection of best-fit models of protein evolution." *Bioinformatics* 21:2104-2105. doi: 10.1093/bioinformatics/bti263.

Adams, Mark D., Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Scott N. Henderson, Granger G. Sutton, Jennifer R. Wortman, Mark D. Yandell, Qing Zhang, Lin X. Chen, Rhonda C. Brandon, Yu-Hui C. Rogers, Robert G. Blazej, Mark Champe, Barret D. Pfeiffer, Kenneth H. Wan, Clare Doyle, Evan G. Baxter, Gregg Helt, Catherine R. Nelson, George L. Gabor, Miklos, Josep F. Abril, Anna Agbayani, Hui-Jin An, Cynthia Andrews-Pfannkoch, Danita Baldwin, Richard M. Ballew, Anand Basu, James Baxendale, Leyla Bayraktaroglu, Ellen M. Beasley, Karen Y. Beeson, P. V. Benos, Benjamin P. Berman, Deepali Bhandari, Slava Bolshakov, Dana Borkova, Michael R. Botchan, John Bouck, Peter Brokstein, Phillipe Brottier, Kenneth C. Burtis, Dana A. Busam, Heather Butler, Edouard Cadieu, Angela Center, Ishwar Chandra, J. Michael Cherry, Simon Cawley, Carl Dahlke, Lionel B. Davenport, Peter Davies, Beatriz de Pablos, Arthur Delcher, Zuoming Deng, Anne Deslattes Mays, Ian Dew, Suzanne M. Dietz, Kristina Dodson, Lisa E. Doup, Michael Downes, Shannon Dugan-Rocha, Boris C. Dunkov, Patrick Dunn, Kenneth J. Durbin, Carlos C. Evangelista, Concepcion Ferraz, Steven Ferriera, Wolfgang Fleischmann, Carl Fosler, Andrei E. Gabrielian, Neha S. Garg, William M. Gelbart, Ken Glasser, Anna Glodek, Fangcheng Gong, J. Harley Gorrell, Zhiping Gu, Ping Guan, Michael Harris, Nomi L. Harris, Damon Harvey, Thomas J. Heiman, Judith R. Hernandez, Jarrett Houck, Damon Hostin, Kathryn A. Houston, Timothy J. Howland, Ming-Hui Wei, Chinyere Ibegwam, Mena Jalali, Francis Kalush, Gary H. Karpen, Zhaoxi Ke, James A. Kennison, Karen A. Ketchum, Bruce E. Kimmel, Chinnappa D. Kodira, Cheryl Kraft, Saul Kravitz, David Kulp, Zhongwu Lai, Paul Lasko, Yiding Lei, Alexander A. Levitsky, Jiayin Li, Zhenya Li, Yong Liang, Xiaoying Lin, Xiangjun Liu, Bettina Mattei, Tina C. McIntosh, Michael P. McLeod, Duncan McPherson, Gennady Merkulov, Natalia V. Milshina, Clark Mobarry, Joe Morris, Ali Moshrefi, Stephen M. Mount, Mee Moy, Brian Murphy, Lee Murphy, Donna M. Muzny, David L. Nelson, David R. Nelson, Keith A. Nelson, Katherine Nixon, Deborah R. Nusskern, Joanne M. Pacleb, Michael Palazzolo, Gjange S. Pittman, Sue Pan, John Pollard, Vinita Puri, Martin G. Reese, Knut Reinert, Karin Remington, Robert D. C. Saunders, Frederick Scheeler, Hua Shen, Bixiang Christopher Shue, Inga Sidén-Kiamos, Michael Simpson, Marian P. Skupski, Tom Smith, Eugene Spier, Allan C. Spradling, Mark Stapleton, Renee Strong, Eric Sun, Robert Svirskas, Cyndee Tector, Russell Turner, Eli Venter, Aihui H. Wang, Xin Wang, Zhen-Yuan Wang, David A. Wassarman, George M. Weinstock, Jean Weissenbach, Sherita M. Williams, Trevor Woodage, Kim C. Worley, David Wu, Song Yang, Q. Alison Yao, Jane Ye, Ru-Fang Yeh, Jayshree S. Zaveri, Ming Zhan, Guangren Zhang, Qi Zhao, Liansheng Zheng, Xiangqun H. Zheng, Fei N. Zhong, Wenyan Zhong, Xiaojun Zhou, Shiaoping Zhu, Xiaohong Zhu, Hamilton O. Smith, Richard A. Gibbs, Eugene W. Myers, Gerald M. Rubin, and J. Craig Venter. 2000. "The Genome Sequence of Drosophila melanogaster." *Science* 287:2185-2195. doi: 10.1126/science.287.5461.2185.

Adams, Melanie A., Michael D. L. Suits, Jimin Zheng, and Zongchao Jia. 2007. "Piecing together the structure–function puzzle: Experiences in structure-based functional annotation of hypothetical proteins." *PROTEOMICS* 7:2920-2932. doi: 10.1002/pmic.200700099.

Adebali, Ogun, and Igor B. Zhulin. 2017. "Aquerium: a web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree." *Proteins* 85:72-77. doi: 10.1002/prot.25199.

Agnew, Philip, JJ Becnel, Dieter Ebert, and Y Michalakis. 2003. "Symbiosis of microsporidia and insects." *Insect Symbiosis. Volume*:145-164.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002a. "How Cells Obtain Energy from Food." In *Molecular Biology of the Cell*. New York: Garland Science.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002b. "Studying Gene Expression and Function." In *Molecular Biology of the Cell*. New York: Garland Science.

Altenhoff, Adrian M, Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Cécile Pereira, Leszek P Pryszcz, Fabian Schreiber, Alan Sousa da Silva, Damian Szklarczyk, Clément-Marie Train, Peer Bork, Odile Lecompte, Christian von Mering, Ioannis Xenarios, Kimmen Sjölander, Lars Juhl Jensen, Maria J Martin, Matthieu Muffato, Adrian M Altenhoff, Brigitte Boeckmann, Salvador Capella-Gutierrez, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Cécile Pereira, Leszek P Pryszcz, Fabian Schreiber, Alan Sousa da Silva, Damian Szklarczyk, Clément-Marie Train, Odile Lecompte, Ioannis Xenarios, Kimmen Sjölander, Maria J Martin, Matthieu Muffato, Toni Gabaldón, Suzanna E Lewis, Paul D Thomas, Erik Sonnhammer, Christophe Dessimoz, Toni Gabaldón, Suzanna E Lewis, Paul D Thomas, Erik Sonnhammer, and Christophe Dessimoz. 2016. "Standardized benchmarking in the quest for orthologs." *Nature Methods* 13:425-430. doi: 10.1038/nmeth.3830.

Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. 2012. "Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs." *PLoS Computational Biology* 8:e1002514. doi: 10.1371/journal.pcbi.1002514.

Altenhoff, Adrian M., Nives Šunca, Natasha Glover, Clément Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, Bartlomiej Tomiczek, Steven Müller, Henning Redestig, Gaston H. Gonnet, and Christophe Dessimoz. 2015. "The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements." *Nucleic Acids Research* 43:D240-D249. doi: 10.1093/nar/gku1158.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Research* 25:3389-3402.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic local alignment search tool." *Journal of Molecular Biology* 215:403-410. doi: 10.1016/S0022-2836(05)80360-2.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. "Gene ontology: Tool for the unification of biology." *Nature Genetics* 25:25-29. doi: 10.1038/75556.

Aurrecoechea, Cristina, Ana Barreto, John Brestelli, Brian P Brunk, Elisabet V Caler, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S Harb, Mark Heiges, John Iodice, Jessica C Kissinger, Eileen T Kraemer, Wei Li, Vishal Nayak, Cary Pennington, Deborah F Pinney, Brian Pitts, David S Roos, Ganesh Srinivasamoorthy, Christian J Stoeckert, Charles Treatman, and Haiming Wang. 2011. "AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species." *Nucleic acids research* 39:D612-9. doi: 10.1093/nar/gkq1006.

Baker, D. 2001. "Protein Structure Prediction and Structural Genomics." *Science* 294:93-96. doi: 10.1126/science.1065659.

Bakowski, Malina A., Margaret Priest, Sarah Young, Christina A. Cuomo, and Emily R. Troemel. 2014. "Genome Sequence of the Microsporidian Species Nematocida sp1 Strain ERTm6 (ATCC PRA-372)." *Genome Announcements* 2:e00905-14. doi: 10.1128/genomeA.00905-14.

Balbiani, G. 1882. "Sur les microsporidies ou psorospermies des Articulés." *C. R. Acad. Sci.* 95:1168–1171.

Bargsten, Joachim W., Edouard I. Severing, Jan-Peter Nap, Gabino F. Sanchez-Perez, and Aalt D.J. van Dijk. 2014. "Biological process annotation of proteins across the plant kingdom." *Current Plant Biology* 1:73-82. doi: 10.1016/j.cpb.2014.07.001.

Bateman, Alex, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang. 2017. "UniProt: the universal protein knowledgebase." *Nucleic Acids Research* 45:D158-D169. doi: 10.1093/nar/gkw1099.

Baum, David A., Stacey DeWitt Smith, and Samuel S. S. Donovan. 2005. "The Tree-Thinking Challenge." *Science* 310:979-980. doi: 10.1126/science.1117727.

Belkorchia, Abdel, Corinne Biderre, Cécile Militon, Valérie Polonais, Patrick Wincker, Claire Jubin, Frédéric Delbac, Eric Peyretaillade, and Pierre Peyret. 2008. "In vitro propagation of the microsporidian pathogen Brachiola algerae and studies of its chromosome and ribosomal DNA organization in the context of the complete genome sequencing project." *Parasitology International* 57:62-71. doi: 10.1016/j.parint.2007.09.002.

Bergsten, Johannes. 2005. "A review of long-branch attraction." *Cladistics* 21:163-193. doi: 10.1111/j.1096-0031.2005.00059.x.

Bjørnson, Susan, and David Oi. 2014. "Microsporidia Biological Control Agents and Pathogens of Beneficial Insects." In *Microsporidia*, edited by Louis M. Weiss and James J. Becnel, 635-670. Chichester, UK: John Wiley & Sons, Inc.

Bohne, Wolfgang, Karin Böttcher, and Uwe Groß. 2011. "The parasitophorous vacuole of Encephalitozoon cuniculi: Biogenesis and characteristics of the host cell–pathogen interface." *International Journal of Medical Microbiology* 301:395-399. doi: 10.1016/j.ijmm.2011.04.006.

Bretagne, S., F. Foulet, W. Alkassoum, J. Fleury-Feith, and M. Develoux. 1993. "Prevalence of Enterocytozoon bieneusi spores in the stool of AIDS patients and African children not infected by HIV." *Bulletin De La Societe De Pathologie Exotique (1990)* 86:351-357.

Brown, J. R., and W. F. Doolittle. 1995. "Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications." *Proceedings of the National Academy of Sciences* 92:2441-2445. doi: 10.1073/pnas.92.7.2441.

Canning, Elizabeth U. 1986. *The microsporidia of vertebrates*: Academic Press.

Cantarel, Brandi L., Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes." *Genome Research* 18:188-196. doi: 10.1101/gr.6743907.

Capella-Gutiérrez, Salvador, Marina Marcet-Houben, and Toni Gabaldón. 2012. "Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi." *BMC biology* 10:47-47. doi: 10.1186/1741-7007-10-47.

Capra, John A., Maureen Stolzer, Dannie Durand, and Katherine S. Pollard. 2013. "How old is my gene?" *Trends in Genetics* 29:659-668. doi: 10.1016/j.tig.2013.07.001.

Cavalier-Smith, T. 1983. "A 6-kingdom classification and a unified phylogeny." In *Endocytobiology II: intracellular space as oligogenetic*, edited by HEA.; Schwemmler Schenk, WS., 1027–1034. Berlin: Walter de Gruyter & Co.

Chamberlain, Scott, and Eduard Szocs. 2013. "taxize - taxonomic search and retrieval in R." *F1000Research*.

Chamberlain, Scott, Eduard Szoecs, Zachary Foster, Zebulun Arendsee, Carl Boettiger, Karthik Ram, Ignasi Bartomeus, John Baumgartner, James O'Donnell, Jari Oksanen, Bastian Greshake Tzovaras, Philippe

Marchand, and Ngoc-Vinh Tran. 2018. *taxize: Taxonomic information from around the web*.

Charbonneau, Lise R., Neil Kirk Hillier, Richard E. L. Rogers, Geoffrey R. Williams, and Dave Shutler. 2016. "Effects of Nosema apis, N. ceranae, and coinfections on honey bee (Apis mellifera) learning and memory." *Scientific Reports* 6. doi: 10.1038/srep22626.

Chen, Xiaoshu, and Jianzhi Zhang. 2012. "The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data." *PLoS Computational Biology* 8:e1002784. doi: 10.1371/journal.pcbi.1002784.

Cheng, Hui-Wen A., Frances E. Lucy, Thaddeus K. Graczyk, Michael A. Broaders, and Sergey E. Mastitsky. 2011. "Municipal wastewater treatment plants as removal systems and environmental sources of human-virulent microsporidian spores." *Parasitology Research* 109:595-603. doi: 10.1007/s00436-011-2291-x.

Chothia, C, and A M Lesk. 1986. "The relation between the divergence of sequence and structure in proteins." *The EMBO Journal* 5:823-826.

Choudhary, Maria M., Maureen G. Metcalfe, Kathryn Arrambide, Caryn Bern, Govinda S. Visvesvara, Norman J. Pieniazek, Rebecca D. Bandea, Marlene DeLeon-Carnes, Patricia Adem, Moaz M. Choudhary, Sherif R. Zaki, and Musab U. Saeed. 2011. "Tubulinosema sp. Microsporidian Myositis in Immunosuppressed Patient." *Emerging Infectious Diseases* 17:1727-1730. doi: 10.3201/eid1709.101926.

Choudhuri, Supratim. 2014. "Phylogenetic Analysis." In *Bioinformatics for Beginners*, 209-218. Oxford: Academic Press.

Corradi, Nicolas, and Patrick J. Keeling. 2009. "Microsporidia: a journey through radical taxonomical revisions." *Fungal Biology Reviews* 23:1-8. doi: 10.1016/j.fbr.2009.05.001.

Corradi, Nicolas, Jean-François Pombert, Laurent Farinelli, Elizabeth S. Didier, and Patrick J. Keeling. 2010. "The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis." *Nature Communications* 1:77. doi: 10.1038/ncomms1082.

Coyle, Christina M., Louis M. Weiss, Luther V. Rhodes, Ann Cali, Peter M. Takvorian, Daniel F. Brown, Govinda S. Visvesvara, Lihua Xiao, Jaan Naktin, Eric Young, Marcelo Gareca, Georgia Colasante, and Murray Wittner. 2004. "Fatal Myositis Due to the Microsporidian Brachiola algerae, a Mosquito Pathogen." *The New England journal of medicine* 351:42-47. doi: 10.1056/NEJMoa032655.

Cuomo, Christina A., Christopher A. Desjardins, Malina A. Bakowski, Jonathan Goldberg, Amy T. Ma, James J. Becnel, Elizabeth S. Didier, Lin Fan, David I. Heiman, Joshua Z. Levin, Sarah Young, Qiandong Zeng, and Emily R. Troemel. 2012. "Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth." *Genome Research* 22:2478-2488. doi: 10.1101/gr.142802.112.

Date, Shailesh V., and José M. Peregrín-Alvarez. 2008. "Phylogenetic profiling." *Methods in Molecular Biology* 453:201-216. doi: 10.1007/978-1-60327-429-6-9.

Daubin, Vincent, Manolo Gouy, and Guy Perrière. 2002. "A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history." *Genome Research* 12:1080-1090. doi: 10.1101/gr.187002.

Dean, Paul, Robert P. Hirt, and T. Martin Embley. 2016. "Microsporidia: Why Make Nucleotides if You Can Steal Them?" *PLoS Pathogens* 12. doi: 10.1371/journal.ppat.1005870.

Decraene, V., M. Lebbad, S. Botero-Kleiven, A.-M. Gustavsson, and M. Löfdahl. 2012. "First reported foodborne outbreak associated with microsporidia, Sweden, October 2009." *Epidemiology and Infection* 140:519-527. doi: 10.1017/S095026881100077X.

Desjardins, Christopher A., Neil D. Sanscrainte, Jonathan M. Goldberg, David Heiman, Sarah Young, Qiandong Zeng, Hiten D. Madhani, James J. Becnel, and Christina A. Cuomo. 2015. "Contrasting host–pathogen interactions and genome evolution in two generalist and specialist microsporidian pathogens of mosquitoes." *Nature Communications* 6:7121. doi: 10.1038/ncomms8121.

Desportes, I., Y. Le Charpentier, A. Galian, F. Bernard, B. Cochand-Priollet, A. Lavergne, P. Ravisse, and R. Modigliani. 1985. "Occurrence of a new microsporidan: Enterocytozoon bieneusi n.g., n. sp., in the enterocytes of a human patient with AIDS." *The Journal of Protozoology* 32:250-254.

Dey, Gautam, Ariel Jaimovich, Sean R. Collins, Akiko Seki, and Tobias Meyer. 2015. "Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling." *Cell Reports* 10:993-1006. doi: 10.1016/j.celrep.2015.01.025.

Didier, Elizabeth S., and Louis M. Weiss. 2011. "Microsporidiosis: Not just in AIDS patients." *Current opinion in infectious diseases* 24:490-495. doi: 10.1097/QCO.0b013e32834aa152.

Dolgikh, Viacheslav V. 2000. "Activities of enzymes of carbohydrate and energy metabolism of the intracellular stages of the microsporidian, Nosema grylli." *Protistology* 1:87-91.

Dolgikh, Viacheslav V., Julia J. Sokolova, and Irma V. Issi. 1997. "Activities of enzymes of carbohydrate and energy metabolism of the spores of the microsporidian, Nosema grylli." *Journal of Eukaryotic Microbiology* 44:246-249. doi: 10.1111/j.1550-7408.1997.tb05707.x.

Ebersberger, I., A. von Haeseler, and HA. Schmidt. 2007. "Phylogeny Reconstruction." In *Bioinformatics: From Genomes to Therapies*, edited by T. Lengauer, 83-128. Weinheim: Wiley-VCH.

Ebersberger, Ingo, Stefan Simm, Matthias S. Leisegang, Peter Schmitzberger, Oliver Mirus, Arndt von Haeseler, Markus T. Bohnsack, and Enrico Schleiff. 2014. "The evolution of the ribosome biogenesis pathway from a yeast perspective." *Nucleic Acids Research* 42:1509-1523. doi: 10.1093/nar/gkt1137.

Ebersberger, Ingo, Sascha Strauss, and Arndt von Haeseler. 2009. "HaMStR: profile hidden markov model based search for orthologs in ESTs." *BMC evolutionary biology* 9:157-157. doi: 10.1186/1471-2148-9-157.

Eddy, S. R. 1998. "Profile hidden Markov models." *Bioinformatics (Oxford, England)* 14:755-763.

Eddy, Sean R. 2004. "Where did the BLOSUM62 alignment score matrix come from?" *Nature Biotechnology* 22:1035-1036. doi: 10.1038/nbt0804-1035.

Edlind, Thomas D, Jing Li, Govinda S Visvesvara, Michael H Vodkin, Gerald L McLaughlin, and Santosh K Katiyar. 1996. "Phylogenetic Analysis of β-Tubulin Sequences from Amitochondrial Protozoa." *Molecular Phylogenetics and Evolution* 5:359-367. doi: 10.1006/mpev.1996.0031.

Edwards, A W F. 1996. "The Origin and Early Development of the Method of Minimum Evolution for the Reconstruction of …." *Systematic Biology*.

Fast, N M, and P J Keeling. 2001. "Alpha and beta subunits of pyruvate dehydrogenase E1 from the microsporidian Nosema locustae: mitochondrion-derived carbon metabolism in microsporidia." *Molecular and biochemical parasitology* 117:201-9.

Federhen, Scott. 2012. "The NCBI Taxonomy." *Nucleic Acids Res.* 40:D136-D143. doi: 10.1093/nar/gkr1178.

Felsenstein, Joseph. 1978. "Cases in which Parsimony or Compatibility Methods Will be Positively Misleading." *Systematic Zoology* 27:401-410. doi: 10.2307/2412923.

Felsenstein, Joseph. 1985. "Confidence Limits on Phylogenies: An Approach Using the Bootstrap." *Evolution* 39:783. doi: 10.2307/2408678.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L.L. Sonnhammer, John Tate, and Marco Punta. 2014. "Pfam: The protein families database." *Nucleic Acids Research* 42. doi: 10.1093/nar/gkt1223.

Finn, Robert D., Jody Clements, William Arndt, Benjamin L. Miller, Travis J. Wheeler, Fabian Schreiber, Alex Bateman, and Sean R. Eddy. 2015. "HMMER web server: 2015 update." *Nucleic Acids Research* 43:W30-W38. doi: 10.1093/nar/gkv397.

Finn, Robert D., Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. 2016. "The Pfam protein families database: towards a more sustainable future." *Nucleic Acids Research* 44:D279-D285. doi: 10.1093/nar/gkv1344.

Fitch, Walter M. 1970. "Distinguishing Homologous from Analogous Proteins." *Systematic Zoology* 19:99. doi: 10.2307/2412448.

Fourment, Mathieu, and Mark J Gibbs. 2006. "PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change." *BMC Evolutionary Biology* 6:1. doi: 10.1186/1471-2148-6-1.

Friedberg, Iddo. 2006. "Automated protein function prediction—the genomic challenge." *Briefings in Bioinformatics* 7:225-242. doi: 10.1093/bib/bbl004.

Futuyma, Douglas J. 2005. *Evolution*: Sinauer Associates Inc.

Gabaldón, T., and M. A. Huynen. 2004. "Prediction of protein function and pathways in the genome era." *Cellular and Molecular Life Sciences (CMLS)* 61:930-944. doi: 10.1007/s00018-003-3387-y.

Gabaldón, Toni. 2007. "Evolution of proteins and proteomes: a phylogenetics approach." *Evolutionary Bioinformatics Online* 1:51-61.

Gabaldón, Toni. 2008. "Large-scale assignment of orthology: back to phylogenetics?" *Genome Biology* 9:235. doi: 10.1186/gb-2008-9-10-235.

Gabaldón, Toni, and Eugene V. Koonin. 2013. "Functional and evolutionary implications of gene orthology." *Nature Reviews Genetics* 14:360-366. doi: 10.1038/nrg3456.

Garcia, Lynne S. 2002. "Laboratory Identification of the Microsporidia." *Journal of Clinical Microbiology* 40:1892-1901. doi: 10.1128/JCM.40.6.1892-1901.2002.

Gaucher, Eric A., James T. Kratzer, and Ryan N. Randall. 2010. "Deep Phylogeny—How a Tree Can Help Characterize Early Life on Earth." *Cold Spring Harbor Perspectives in Biology* 2. doi: 10.1101/cshperspect.a002238.

Germot, Agnes, Hervé Philippe, and Hervé Le Guyader. 1997. "Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae." *Molecular and Biochemical Parasitology*:10.

Germot, Agnès, Hervé Philippe, and Hervé Le Guyader. 1996. "Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 93:14614-14617.

Gill, Erin E., and Naomi M. Fast. 2006. "Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes." *Gene* 375:103-109. doi: 10.1016/j.gene.2006.02.023.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. 1996. "Life with 6000 Genes." *Science* 274:546-567. doi: 10.1126/science.274.5287.546.

Götz, Stefan, Juan Miguel García-Gómez, Javier Terol, Tim D. Williams, Shivashankar H. Nagaraj, María José Nueda, Montserrat Robles, Manuel Talón, Joaquín Dopazo, and Ana Conesa. 2008. "High-throughput functional annotation and data mining with the Blast2GO suite." *Nucleic Acids Research* 36:3420-3435. doi: 10.1093/nar/gkn176.

Gregory, T. Ryan. 2008. "Understanding Evolutionary Trees." *Evolution: Education and Outreach* 1:121-137. doi: 10.1007/s12052-008-0035-x.

Grimm, David. 2006. "A Mouse for Every Gene." *Science* 312:1862-1866. doi: 10.1126/science.312.5782.1862.

Hall, Bradford, Advait Limaye, and Ashok B Kulkarni. 2009. "Overview: Generation of Gene Knockout Mice." *Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.]* CHAPTER:Unit-19.1217. doi: 10.1002/0471143030.cb1912s44.

Heinz, Eva, Christian Hacker, Paul Dean, John Mifsud, Alina V. Goldberg, Tom A. Williams, Sirintra Nakjang, Alison Gregory, Robert P. Hirt, John M. Lucocq, Edmund R.S. Kunji, and T. Martin Embley. 2014. "Plasma Membrane-Located Purine Nucleotide Transport Proteins Are Key Components for Host Exploitation by Microsporidian Intracellular Parasites." *PLoS Pathogens* 10. doi: 10.1371/journal.ppat.1004547.

Heinz, Eva, Tom a Williams, Sirintra Nakjang, Christophe J Noël, Daniel C Swan, Alina V Goldberg, Simon R Harris, Thomas Weinmaier, Stephanie Markert, Dörte Becher, Jörg Bernhardt, Tal Dagan, Christian Hacker, John M Lucocq, Thomas Schweder, Thomas Rattei, Neil Hall, Robert P Hirt, and T Martin Embley. 2012. "The genome of the obligate intracellular parasite Trachipleistophora hominis: new insights into microsporidian genome dynamics and reductive evolution." *PLoS pathogens* 8:e1002979-e1002979. doi: 10.1371/journal.ppat.1002979.

Hibbett, David S., Manfred Binder, Joseph F. Bischoff, Meredith Blackwell, Paul F. Cannon, Ove E. Eriksson, Sabine Huhndorf, Timothy James, Paul M. Kirk, Robert Lücking, H. Thorsten Lumbsch, François Lutzoni, P. Brandon Matheny, David J. McLaughlin, Martha J. Powell, Scott Redhead, Conrad L. Schoch, Joseph W. Spatafora, Joost A. Stalpers, Rytas Vilgalys, M. Catherine Aime, André Aptroot, Robert Bauer, Dominik Begerow, Gerald L. Benny, Lisa A. Castlebury, Pedro W. Crous, Yu-Cheng Dai, Walter Gams, David M. Geiser, Gareth W. Griffith, Cécile Gueidan, David L. Hawksworth, Geir Hestmark, Kentaro Hosaka, Richard A. Humber, Kevin D. Hyde, Joseph E. Ironside, Urmas Kõljalg, Cletus P. Kurtzman, Karl-Henrik Larsson, Robert Lichtwardt, Joyce Longcore, Jolanta Miądlikowska, Andrew Miller, Jean-Marc Moncalvo, Sharon Mozley-Standridge, Franz Oberwinkler, Erast Parmasto, Valérie Reeb, Jack D. Rogers, Claude Roux, Leif Ryvarden, José Paulo Sampaio, Arthur Schüßler, Junta Sugiyama, R. Greg Thorn, Leif Tibell, Wendy A. Untereiner, Christopher Walker, Zheng Wang, Alex Weir, Michael

Weiss, Merlin M. White, Katarina Winka, Yi-Jian Yao, and Ning Zhang. 2007. "A higher-level phylogenetic classification of the Fungi." *Mycological Research* 111:509-547. doi: 10.1016/j.mycres.2007.03.004.

Hinchliff, Cody E., Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. 2015. "Synthesis of phylogeny and taxonomy into a comprehensive tree of life." *Proceedings of the National Academy of Sciences* 112:12764-12769. doi: 10.1073/pnas.1423041112.

Hirt, R. P., J. M. Logsdon, 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." *Proceedings of the National Academy of Sciences* 96:580-585. doi: 10.1073/pnas.96.2.580.

Hirt, Robert P., Bryan Healy, Charles R. Vossbrinck, Elizabeth U. Canning, and T. Martin Embley. 1997. "A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria." *Current Biology* 7:995-998. doi: 10.1016/S0960-9822(06)00420-9.

Hirt, Robert P., and David S. Horner. 2004. *Organelles, Genomes and Eukaryote Phylogeny: An Evolutionary Synthesis in the Age of Genomics*: CRC Press.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." *Molecular Biology and Evolution* 33:1635-1638. doi: 10.1093/molbev/msw046.

Jain, Arpit, Arndt von Haeseler, and Ingo Ebersberger. 2018. "The evolutionary traceability of proteins." *bioRxiv*:302109. doi: 10.1101/302109.

James, Timothy Y, Adrian Pelin, Linda Bonen, Steven Ahrendt, Divya Sain, Nicolas Corradi, and Jason E Stajich. 2013. "Shared signatures of parasitism and phylogenomics unite Cryptomycota and microsporidia." *Current biology : CB* 23:1548-53. doi: 10.1016/j.cub.2013.06.057.

James, Timothy Y., Frank Kauff, Conrad L. Schoch, P. Brandon Matheny, Valérie Hofstetter, Cymon J. Cox, Gail Celio, Cécile Gueidan, Emily Fraker, Jolanta Miadlikowska, H. Thorsten Lumbsch, Alexandra Rauhut, Valérie Reeb, A. Elizabeth Arnold, Anja Amtoft, Jason E. Stajich, Kentaro Hosaka, Gi-Ho Sung, Desiree Johnson, Ben O'Rourke, Michael Crockett, Manfred Binder, Judd M. Curtis, Jason C. Slot, Zheng Wang, Andrew W. Wilson, Arthur Schüßler, Joyce E. Longcore, Kerry O'Donnell, Sharon Mozley-Standridge, David Porter, Peter M. Letcher, Martha J. Powell, John W. Taylor, Merlin M. White, Gareth W. Griffith, David R. Davies, Richard A. Humber, Joseph B. Morton, Junta Sugiyama, Amy Y.

Rossman, Jack D. Rogers, Don H. Pfister, David Hewitt, Karen Hansen, Sarah Hambleton, Robert A. Shoemaker, Jan Kohlmeyer, Brigitte Volkmann-Kohlmeyer, Robert A. Spotts, Maryna Serdani, Pedro W. Crous, Karen W. Hughes, Kenji Matsuura, Ewald Langer, Gitta Langer, Wendy A. Untereiner, Robert Lücking, Burkhard Büdel, David M. Geiser, André Aptroot, Paul Diederich, Imke Schmitt, Matthias Schultz, Rebecca Yahr, David S. Hibbett, François Lutzoni, David J. McLaughlin, Joseph W. Spatafora, and Rytas Vilgalys. 2006. "Reconstructing the early evolution of Fungi using a six-gene phylogeny." *Nature* 443:818-822. doi: 10.1038/nature05110.

Jedrzejewski, Szymon, Thaddeus K. Graczyk, Anna Slodkowicz-Kowalska, Leena Tamang, and Anna C. Majewska. 2007. "Quantitative Assessment of Contamination of Fresh Food Produce of Various Retail Types by Human-Virulent Microsporidian Spores." *Applied and Environmental Microbiology* 73:4071-4073. doi: 10.1128/AEM.00477-07.

Jiří, Vávra, Yachnis Anthony T., Shadduck John A., and Orenstein Jan M. 2007. "Microsporidia of the Genus Trachipleistophora—Causative Agents of Human Microsporidiosis: Description of Trachipleistophora anthropophthera N. Sp. (Protozoa: Microsporidia)." *Journal of Eukaryotic Microbiology* 45:273-283. doi: 10.1111/j.1550-7408.1998.tb04536.x.

Jothi, Raja, Teresa M Przytycka, and L Aravind. 2007. "Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment." *BMC bioinformatics* 8:173-173. doi: 10.1186/1471-2105-8-173.

Kachroo, Aashiq H., Jon M. Laurent, Christopher M. Yellman, Austin G. Meyer, Claus O. Wilke, and Edward M. Marcotte. 2015. "Systematic humanization of yeast genes reveals conserved functions and genetic modularity." *Science (New York, N.Y.)* 348:921-925. doi: 10.1126/science.aaa0769.

Kamaishi, Takashi, Tetsuo Hashimoto, Yoshihiro Nakamura, Yutaka Masuda, Fuminori Nakamura, Ken-ichi Okamoto, Makoto Shimizu, and Masami Hasegawa. 1996. "Complete Nucleotide Sequences of the Genes Encoding Translation Elongation Factors 1α and 2 from a microsporidian parasite, Glugea plecoglossi: Implications for the Deepest Branching of Eukaryotes." *The Journal of Biochemistry* 120:1095-1103.

Kanehisa, M, and S Goto. 2000. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* 28:27-30.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. "Data, information, knowledge and principle: Back to metabolism in KEGG." *Nucleic Acids Research* 42. doi: 10.1093/nar/gkt1076.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. "KEGG as a reference resource for gene and protein annotation." *Nucleic Acids Research* 44:D457-D462. doi: 10.1093/nar/gkv1070.

Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." *Journal of Molecular Biology* 428:726-731. doi: 10.1016/j.jmb.2015.11.006.

Katinka, M D, S Duprat, E Cornillot, G Méténier, 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 Vivarès. 2001. "Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi." *Nature* 414:450-453. doi: 10.1038/35106579.

Kaya, Ghosh, and Weiss Louis M. 2012. "T cell response and persistence of the microsporidia." *FEMS Microbiology Reviews* 36:748-760. doi: 10.1111/j.1574-6976.2011.00318.x.

Keeling, P. J., and W. F. Doolittle. 1996. "Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family." *Molecular Biology and Evolution* 13:1297-1305. doi: 10.1093/oxfordjournals.molbev.a025576.

Keeling, Patrick. 2009. "Five questions about microsporidia." *PLoS pathogens* 5:e1000489-e1000489. doi: 10.1371/journal.ppat.1000489.

Keeling, Patrick J. 2003. "Congruent evidence from α-tubulin and β-tubulin gene phylogenies for a zygomycete origin of microsporidia." *Fungal Genetics and Biology* 38:298-309. doi: 10.1016/S1087-1845(02)00537-6.

Keeling, Patrick J, and Nicolas Corradi. 2011. "Shrink it or lose it: balancing loss of function with shrinking genomes in the microsporidia." *Virulence* 2:67-70. doi: 10.4161/viru.2.1.14606.

Keeling, Patrick J, and Naomi M Fast. 2002. "Microsporidia: biology and evolution of highly reduced intracellular parasites." *Annual review of microbiology* 56:93-116. doi: 10.1146/annurev.micro.56.012302.160854.

Keeling, Patrick J., Melissa A. Luker, and Jeffrey D. Palmer. 2000. "Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi." *Molecular Biology and Evolution* 17:23-31. doi: 10.1093/oxfordjournals.molbev.a026235.

Kensche, Philip R, Vera van Noort, Bas E Dutilh, and Martijn A Huynen. 2008. "Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution." *Journal of the Royal Society, Interface / the Royal Society* 5:151-70. doi: 10.1098/rsif.2007.1047.

Kishino, Hirohisa, and Masami Hasegawa. 1989. "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea." *Journal of Molecular Evolution* 29:170-179. doi: 10.1007/BF02100115.

Kmmari, Suresh, Srinu Rathlavath, Devika Pillai, and Gadasu Rajesh. 2018. "Hepatopancreatic Microsporidiasis (HPM) in Shrimp Culture: A

Review." *International Journal of Current Microbiology and Applied Sciences* 7:3208-3215. doi: 10.20546/ijcmas.2018.701.383.

Koestler, Tina, and Ingo Ebersberger. 2011. "Zygomycetes, Microsporidia, and the Evolutionary Ancestry of Sex Determination." *Genome Biology and Evolution* 3:186-194. doi: 10.1093/gbe/evr009.

Koestler, Tina, Arndt von Haeseler, and Ingo Ebersberger. 2010. "FACT: functional annotation transfer between proteins with similar feature architectures." *BMC bioinformatics* 11:417-417. doi: 10.1186/1471-2105-11-417.

Koonin, Eugene V. 2005. "Orthologs, Paralogs, and Evolutionary Genomics." *Annual Review of Genetics* 39:309-338. doi: 10.1146/annurev.genet.39.073003.114725.

Krieg, A. 1955. "Ueber Infektionskrankheiten bei Engerlingen von Melolontha spec. unter besonderer Beruecksichtigung einer Mikrosporidien-Erkrankung." *Zentr. Bakteriol. Parasitenk* 108:535-538.

Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin. 2011. "Computational methods for Gene Orthology inference." *Briefings in Bioinformatics* 12:379-391. doi: 10.1093/bib/bbr030.

Kück, Patrick, Christoph Mayer, Johann-Wolfgang Wägele, and Bernhard Misof. 2012. "Long Branch Effects Distort Maximum Likelihood Phylogenies in Simulations Despite Selection of the Correct Model." *PLoS ONE* 7:e36593. doi: 10.1371/journal.pone.0036593.

Kudo, R. R., and E. W. Daniels. 1963. "An Electron Microscope Study of the Spore of a Microsporidian, Thelohania californica*." *The Journal of Protozoology* 10:112-120. doi: 10.1111/j.1550-7408.1963.tb01645.x.

Kupczok, Anne, Heiko A Schmidt, and Arndt von Haeseler. 2010. "Accuracy of phylogeny reconstruction methods combining overlapping gene data sets." *Algorithms for Molecular Biology : AMB* 5:37. doi: 10.1186/1748-7188-5-37.

Lan, Ning, R. Jansen, and M. Gerstein. 2002. "Toward a systematic definition of protein function that scales to the genome level: defining function in terms of interactions." *Proceedings of the IEEE* 90:1848-1858. doi: 10.1109/JPROC.2002.805302.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson,

Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg,

Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. 2001. "Initial sequencing and analysis of the human genome." *Nature* 409:860-921. doi: 10.1038/35057062.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. "Clustal W and Clustal X version 2.0." *Bioinformatics* 23:2947-2948. doi: 10.1093/bioinformatics/btm404.

Laskowski, Roman A. 2009. "Integrated Servers for Structure-Informed Function Prediction." In *From Protein Structure to Function with Bioinformatics*, 251-272. Springer, Dordrecht.

Le, Si Quang, and Olivier Gascuel. 2008. "An improved general amino acid replacement matrix." *Molecular Biology and Evolution* 25:1307-1320. doi: 10.1093/molbev/msn067.

Lee, David, Oliver Redfern, and Christine Orengo. 2007. "Predicting protein function from sequence and structure." *Nat. Rev. Mol. Cell Biol.* 8:995-1005. doi: 10.1038/nrm2281.

Lee, John Hwa. 2008. "Molecular Detection of Enterocytozoon bieneusi and Identification of a Potentially Human-Pathogenic Genotype in Milk." *Applied and Environmental Microbiology* 74:1664-1666. doi: 10.1128/AEM.02110-07.

Lee, Jooyoung, Sitao Wu, and Yang Zhang. 2009. "Ab Initio Protein Structure Prediction." In *From Protein Structure to Function with Bioinformatics*, 3-25. Springer, Dordrecht.

Lee, Soo Chan, Nicolas Corradi, Edmond J. Byrnes, Santiago Torres-Martinez, Fred S. Dietrich, Patrick J. Keeling, and Joseph Heitman. 2008. "Microsporidia evolved from ancestral sexual fungi." *Current biology : CB* 18:1675-1679. doi: 10.1016/j.cub.2008.09.030.

Letunic, Ivica, and Peer Bork. 2018. "20 years of the SMART protein domain annotation resource." *Nucleic Acids Research* 46:D493-D496. doi: 10.1093/nar/gkx922.

Li, Li, Christian J Stoeckert, and David S Roos. 2003. "OrthoMCL: identification of ortholog groups for eukaryotic genomes." *Genome research* 13:2178-89. doi: 10.1101/gr.1224503.

Li, Teng, Jimeng Hua, April M Wright, Ying Cui, Qiang Xie, Wenjun Bu, and David M Hillis. 2014. "Long-branch attraction and the phylogeny of true water bugs (Hemiptera: Nepomorpha) as estimated from mitochondrial genomes." *BMC Evolutionary Biology* 14:99. doi: 10.1186/1471-2148-14-99.

Li, Yang, Sarah E. Calvo, Roee Gutman, Jun S. Liu, and Vamsi K. Mootha. 2014. "Expansion of Biological Pathways Based on Evolutionary Inference." *Cell* 158:213-225. doi: 10.1016/j.cell.2014.05.034.

Liu, Shu Q. 2007. *Bioregenerative Engineering: Principles and Applications*: John Wiley & Sons.

Loewenstein, Yaniv, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, and Anna Tramontano. 2009. "Protein function annotation by homology-based inference." *Genome Biology* 10:207. doi: 10.1186/gb-2009-10-2-207.

Lores, Beatriz, Isabel Lopez-Miragaya, Cristina Arias, Soledad Fenoy, Julio Torres, and Carmen del Aguila. 2002. "Intestinal Microsporidiosis Due to Enterocytozoon bieneusi in Elderly Human Immunodeficiency Virus–Negative Patients from Vigo, Spain." *Clinical Infectious Diseases* 34:918-921. doi: 10.1086/339205.

Luallen, Robert J, Aaron W Reinke, Linda Tong, Michael R Botts, Marie-Anne Félix, and Emily R Troemel. 2016. "Discovery of a Natural Microsporidian Pathogen with a Broad Tissue Tropism in Caenorhabditis elegans." *PLOS Pathogens*:28.

Major, Peter, T. Martin Embley, and Tom A. Williams. 2017. "Phylogenetic Diversity of NTT Nucleotide Transport Proteins in Free-Living and Parasitic Bacteria and Eukaryotes." *Genome Biology and Evolution* 9:480-487. doi: 10.1093/gbe/evx015.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." *The Annals of Mathematical Statistics* 18:50-60.

Mathis, Alexander, Rainer Weber, and Peter Deplazes. 2005. "Zoonotic Potential of the Microsporidia." *Clinical Microbiology Reviews* 18:423-445. doi: 10.1128/CMR.18.3.423-445.2005.

Matos, Olga, Maria Luisa Lobo, and Lihua Xiao. 2012. "Epidemiology of Enterocytozoon bieneusi Infection in Humans." [Research article], Last Modified 2012.

Méténier, Guy, and Christian P. Vivarès. 2001. "Molecular characteristics and physiology of microsporidia." *Microbes and Infection* 3:407-415. doi: 10.1016/S1286-4579(01)01398-3.

Moore, A. D., A. Held, N. Terrapon, J. Weiner, and E. Bornberg-Bauer. 2014. "DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins." *Bioinformatics* 30:282-283. doi: 10.1093/bioinformatics/btt640.

Moreira, David, and Purificación López-García. 2007. "The Last Common Ancestor of Modern Cells." In *Lectures in Astrobiology*, edited by Muriel

Gargaud, Hervé Martin and Philippe Claeys, 305-317. Berlin, Heidelberg: Springer Berlin Heidelberg.
Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C Yoshizawa, and Minoru Kanehisa. 2007. "KAAS: an automatic genome annotation and pathway reconstruction server." *Nucleic acids research* 35:W182-5. doi: 10.1093/nar/gkm321.
Mungthin, Mathirut, Ravis Suwannasaeng, Tawee Naaglor, Wirote Areekul, and Saovanee Leelayoova. 2001. "Asymptomatic intestinal microsporidiosis in Thai orphans and child-care workers." *Transactions of the Royal Society of Tropical Medicine and Hygiene* 95:304-306. doi: 10.1016/S0035-9203(01)90243-3.
Nadzirin, Nurul, and Mohd Firdaus-Raih. 2012. "Proteins of unknown function in the protein data bank (PDB): An inventory of true uncharacterized proteins and computational tools for their analysis." *International Journal of Molecular Sciences* 13:12761-12772. doi: 10.3390/ijms131012761.
Naegeli, K. 1857. "Über die neue Krankheit der Seidenraupe und verwandte Organismen." *Botanische Zeitung*, 1857, 760-761. Accessed 2018-03-25 20:33:39.
Nakjang, Sirintra, Tom a Williams, Eva Heinz, Andrew K Watson, Peter G Foster, Kacper M Sendra, Sarah E Heaps, Robert P Hirt, and T Martin Embley. 2013. "Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics." *Genome biology and evolution* 5:2285-303. doi: 10.1093/gbe/evt184.
NCBI Resource Coordinators. 2017. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 45:D12-D17. doi: 10.1093/nar/gkw1071.
Neumann, Peter, and Norman L Carreck. 2010. "Honey bee colony losses." *Journal of Apicultural Research* 49:1-6. doi: 10.3896/IBRA.1.49.1.01.
Noether, Gottfried E. 1987. "Sample Size Determination for Some Common Nonparametric Tests." *Journal of the American Statistical Association* 82:645-647. doi: 10.2307/2289477.
Nordberg, Henrik, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V. Grigoriev, and Inna Dubchak. 2014. "The genome portal of the Department of Energy Joint Genome Institute: 2014 updates." *Nucleic Acids Research* 42:D26-D31. doi: 10.1093/nar/gkt1069.
O'Brien, Kevin P, Maido Remm, and Erik L L Sonnhammer. 2005. "Inparanoid: a comprehensive database of eukaryotic orthologs." *Nucleic acids research* 33:D476-80. doi: 10.1093/nar/gki107.
Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Berlin Heidelberg: Springer-Verlag.
Park, Jong, Kevin Karplus, Christian Barrett, Richard Hughey, David Haussler, Tim Hubbard, and Cyrus Chothia. 1998. "Sequence comparisons using

multiple sequences detect three times as many remote homologues as pairwise methods." *Journal of Molecular Biology* 284:1201-1210. doi: 10.1006/jmbi.1998.2221.

Parks, Sarah L., and Nick Goldman. 2014. "Maximum likelihood inference of small trees in the presence of long branches." *Systematic Biology* 63:798-811. doi: 10.1093/sysbio/syu044.

Pasteur, Louis. 1870. *Études sur la maladie des vers à soie : moyen pratique assuré de la combattre et d'en prévenir le retour*: Paris : Gauthier-Villars, successeur de Mallet-Bachelier.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles." *Proceedings of the National Academy of Sciences* 96:4285-4288. doi: 10.1073/pnas.96.8.4285.

Peyretaillade, Eric, Nicolas Parisot, Valérie Polonais, Sébastien Terrat, Jérémie Denonfoux, Eric Dugat-Bony, Ivan Wawrzyniak, Corinne Biderre-Petit, Antoine Mahul, Sébastien Rimour, Olivier Gonçalves, Stéphanie Bornes, Frédéric Delbac, Brigitte Chebance, Simone Duprat, Gaëlle Samson, Michael Katinka, Jean Weissenbach, Patrick Wincker, and Pierre Peyret. 2012. "Annotation of microsporidian genomes using transcriptional signals." *Nature Communications* 3:1137. doi: 10.1038/ncomms2156.

Philippe, H. 2000. "Opinion: long branch attraction and protist phylogeny." *Protist* 151:307-316. doi: 10.1078/S1434-4610(04)70029-2.

Philippe, Hervé, Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, and Frédéric Delsuc. 2005. "Heterotachy and long-branch attraction in phylogenetics." *BMC Evolutionary Biology* 5:50. doi: 10.1186/1471-2148-5-50.

Pombert, Jean-François, Jinshan Xu, David R. Smith, David Heiman, Sarah Young, Christina A. Cuomo, Louis M. Weiss, and Patrick J. Keeling. 2013. "Complete Genome Sequences from Three Genetically Distinct Strains Reveal High Intraspecies Genetic Diversity in the Microsporidian Encephalitozoon cuniculi." *Eukaryotic Cell* 12:503-511. doi: 10.1128/EC.00312-12.

Ramanan, P., and B. S. Pritt. 2014. "Extraintestinal Microsporidiosis." *Journal of Clinical Microbiology* 52:3839-3844. doi: 10.1128/JCM.00971-14.

Ramsay, Jennifer M., Virginia Watral, Carl B. Schreck, and Michael L. Kent. 2009. "Pseudoloma neurophilia (Microsporidia) infections in zebrafish (Danio rerio): effects of stress on survival, growth and reproduction." *Diseases of aquatic organisms* 88:69-84. doi: 10.3354/dao02145.

Rannala, Bruce, and Ziheng Yang. 1996. "Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference." *Journal of Molecular Evolution* 43:304-311. doi: 10.1007/BF02338839.

Reid, Adam James, Corin Yeats, and Christine Anne Orengo. 2007. "Methods of remote homology detection can be combined to increase coverage by

10% in the midnight zone." *Bioinformatics* 23:2353-2360. doi: 10.1093/bioinformatics/btm355.
Reinke, Aaron W., and Emily R. Troemel. 2015. "The Development of Genetic Modification Techniques in Intracellular Parasites and Potential Applications to Microsporidia." *PLOS Pathogens* 11:e1005283. doi: 10.1371/journal.ppat.1005283.
Rogelio, López-Vélez, Turrientes M. Carmen, Garrón Carla, Montilla Pedro, Navajas Raquel, Fenoy Soledad, and Aguila Carmen. 2006. "Microsporidiosis in Travelers with Diarrhea from the Tropics." *Journal of Travel Medicine* 6:223-227. doi: 10.1111/j.1708-8305.1999.tb00522.x.
Roger, Andrew J., and Alastair G.B. Simpson. 2009. "Evolution: Revisiting the Root of the Eukaryote Tree." *Current Biology* 19:R165-R167. doi: 10.1016/j.cub.2008.12.032.
Rost, Burkhard. 1997. "Protein structures sustain evolutionary drift." *Folding and Design* 2:S19-S24. doi: 10.1016/S1359-0278(97)00059-X.
Rost, Burkhard. 2002. "Enzyme Function Less Conserved than Anticipated." *Journal of Molecular Biology* 318:595-608. doi: 10.1016/S0022-2836(02)00016-5.
Roustan, Valentin, Arpit Jain, Markus Teige, Ingo Ebersberger, and Wolfram Weckwerth. 2016. "An evolutionary perspective of AMPK–TOR signaling in the three domains of life." *Journal of Experimental Botany* 67:3897-3907. doi: 10.1093/jxb/erw211.
Ryan, Ja, and Sl Kohler. 2016. "Distribution, prevalence, and pathology of a microsporidian infecting freshwater sculpins." *Diseases of Aquatic Organisms* 118:195-206. doi: 10.3354/dao02974.
Sael, Lee, Meghana Chitale, and Daisuke Kihara. 2012. "Structure- and Sequence-Based Function Prediction for Non-Homologous Proteins." *Journal of Structural and Functional Genomics* 13:111-123. doi: 10.1007/s10969-012-9126-6.
Santín, Mónica, and Ronald Fayer. 2011. "Microsporidiosis: Enterocytozoon bieneusi in domesticated and wild animals." *Research in Veterinary Science* 90:363-371. doi: 10.1016/j.rvsc.2010.07.014.
Scanlon, Mary, Andrew P. Shaw, Cheng J. Zhou, Govinda S. Visvesvara, and Gordon J. Leitch. 2000. "Infection by microsporidia disrupts the host cell cycle." *Journal of Eukaryotic Microbiology* 47:525-531. doi: 10.1111/j.1550-7408.2000.tb00085.x.
Schmidt, H.A., E. Petzold, M. Vingron, and A. von Haeseler. 2003. "Molecular phylogenetics: parallelized parameter estimation and quartet puzzling." *Journal of Parallel and Distributed Computing* 63:719-727. doi: 10.1016/S0743-7315(03)00129-1.
Schmitt, Thomas, David N. Messina, Fabian Schreiber, and Erik L L Sonnhammer. 2011. "Letter to the Editor: SeqXML and orthoXML:

Standards for sequence and orthology information." *Briefings in Bioinformatics* 12:485-488. doi: 10.1093/bib/bbr025.

Shimodaira, H., and M. Hasegawa. 1999. "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference." *Molecular Biology and Evolution* 16:1114-1116. doi: 10.1093/oxfordjournals.molbev.a026201.

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

Shimodaira, Hidetoshi. 2002. "An Approximately Unbiased Test of Phylogenetic Tree Selection." *Systematic Biology* 51:492-508. doi: 10.1080/10635150290069913.

Slamovits, Claudio H, Naomi M Fast, Joyce S Law, and Patrick J Keeling. 2004. "Genome Compaction and Stability in Microsporidian Intracellular Parasites." *Current Biology* 14:891-896. doi: 10.1016/j.cub.2004.04.041.

Soltis, Douglas E., and Pamela S. Soltis. 2003. "The Role of Phylogenetics in Comparative Genetics." *Plant Physiology* 132:1790-1800. doi: 10.1104/pp.103.022509.

Stamatakis, Alexandros. 2014. "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30:1312-1313. doi: 10.1093/bioinformatics/btu033.

Stanke, M., and B. Morgenstern. 2005. "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints." *Nucleic Acids Research* 33:W465-W467. doi: 10.1093/nar/gki458.

Steel, Mike, Daniel Huson, and Peter J Lockhart. 2000. "Invariable Sites Models and Their Use in Phylogeny Reconstruction." *Systematic Biology*:8.

Stentiford, G.D., J.J. Becnel, L.M. Weiss, P.J. Keeling, E.S. Didier, B.A.P. Williams, S. Bjornson, M.L. Kent, M.A. Freeman, M.J.F. Brown, E.R. Troemel, K. Roesel, Y. Sokolova, K.F. Snowden, and L. Solter. 2016. "Microsporidia – Emergent Pathogens in the Global Food Chain." *Trends in parasitology* 32:336-348. doi: 10.1016/j.pt.2015.12.004.

Studer, Romain A., and Marc Robinson-Rechavi. 2009. "How confident can we be that orthologs are similar, but paralogs differ?" *Trends in Genetics* 25:210-216. doi: 10.1016/j.tig.2009.03.004.

Sukumaran, Jeet, and Mark T. Holder. 2010. "DendroPy: a Python library for phylogenetic computing." *Bioinformatics* 26:1569-1571. doi: 10.1093/bioinformatics/btq228.

Szklarczyk, Damian, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. 2015. "STRING v10: protein–protein interaction networks, integrated over the tree of life." *Nucleic Acids Research* 43:D447-D452. doi: 10.1093/nar/gku1003.

Tanabe, Yuuhiko, Makoto M. Watanabe, and Junta Sugiyama. 2002. "Are Microsporidia really related to Fungi?: a reappraisal based on additional gene sequences from basal fungi." *Mycological Research* 106:1380-1391. doi: 10.1017/S095375620200686X.

Thomarat, Fabienne, Christian P. Vivarès, and Manolo 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." *Journal of Molecular Evolution* 59:780-791. doi: 10.1007/s00239-004-2673-0.

Thomas, Paul D., Valerie Wood, Christopher J. Mungall, Suzanna E. Lewis, Judith A. Blake, and on behalf of the Gene Ontology Consortium. 2012. "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." *PLoS Computational Biology* 8:e1002386. doi: 10.1371/journal.pcbi.1002386.

Tian, Weidong, and Jeffrey Skolnick. 2003. "How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?" *Journal of Molecular Biology* 333:863-882. doi: 10.1016/j.jmb.2003.08.057.

Trachana, Kalliopi, Tomas a Larsson, Sean Powell, Wei-Hua Chen, Tobias Doerks, Jean Muller, and Peer Bork. 2011. "Orthology prediction methods: a quality assessment using curated protein families." *BioEssays : news and reviews in molecular, cellular and developmental biology* 33:769-80. doi: 10.1002/bies.201100062.

Train, Clément-Marie, Natasha M. Glover, Gaston H. Gonnet, Adrian M. Altenhoff, and Christophe Dessimoz. 2017. "Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference." *Bioinformatics* 33:i75-i82. doi: 10.1093/bioinformatics/btx229.

Tran, Ngoc-Vinh, Bastian Greshake Tzovaras, and Ingo Ebersberger. 2018. "PhyloProfile: Dynamic visualization and exploration of multi-layered phylogenetic profiles." *Bioinformatics*. doi: 10.1093/bioinformatics/bty225.

Tsaousis, Anastasios D., Edmund R S Kunji, Alina V. Goldberg, John M. Lucocq, Robert P. Hirt, and T. Martin Embley. 2008. "A novel route for ATP acquisition by the remnant mitochondria of Encephalitozoon cuniculi." *Nature* 453:553-556. doi: 10.1038/nature06903.

van Dongen, Stjin. 2000. "Graph clustering by flow simulation." *Graph stimulation by flow clustering* PhD thesis:University of Utrecht-University of Utrecht. doi: 10.1016/j.cosrev.2007.05.001.

Vandermeer, J. W., and T. A. Gochnauer. 1971. "Trehalase activity associated with spores of Nosema apis." *Journal of Invertebrate Pathology* 17:38-41. doi: 10.1016/0022-2011(71)90122-4.

Vavra, J. 1965. "Study by electron microscope of the morphology and development of some Microsporidia." *Comptes rendus hebdomadaires des seances de l'Academie des sciences. Serie D: Sciences naturelles* 261:3467-3470.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K. Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. 2001. "The Sequence of the Human Genome." *Science* 291:1304-1351. doi: 10.1126/science.1058040.

Vivarès, CP, and G Méténier. 2001. "The microsporidian Encephalitozoon." *Bioessays*:194-202.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. 1987. "Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes." *Nature* 326:411-414. doi: 10.1038/326411a0.

Vossbrinck, Charles R., Bettina A. Debrunner-Vossbrinck, and Louis M. Weiss. 2014. "Phylogeny of the Microsporidia." *Microsporidia*. doi: 10.1002/9781118395264.ch6.

Wang, Tim, Haiyan Yu, Nicholas W. Hughes, Bingxu Liu, Arek Kendirli, Klara Klein, Walter W. Chen, Eric S. Lander, and David M. Sabatini. 2017. "Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras." *Cell* 168:890-903.e15. doi: 10.1016/j.cell.2017.01.013.

Watson, James D., and Janet M. Thornton. 2009. "Case Studies: Function Predictions of Structural Genomics Results." In *From Protein Structure to Function with Bioinformatics*, 273-291. Springer, Dordrecht.

Webb, Edwin C. 1990. "Enzyme Nomenclature." In *The Terminology of Biotechnology: A Multidisciplinary Problem*, 51-60. Springer, Berlin, Heidelberg.

Weiser, Jaroslav. 1964. "On the taxonomic position of the genus Encephalitozoon." *Parasitology* 54:749-751. doi: 10.1017/S0031182000082755.

Weiser, Jaroslav. 1976. "Microsporidia in Invertebrates: Host-Parasite Relations at the Organismal Level." In *Biology of the Microsporidia*, 163-201. Springer, Boston, MA.

Weiss, Louis M., and James J. Becnel. 2014. *Microsporidia: Pathogens of Opportunity*: John Wiley & Sons.

Wetterstrand, KA. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." Available at www.genome.gov/sequencingcostsdata.

Whisstock, James C., and Arthur M. Lesk. 2003. "Prediction of protein function from protein sequence and structure." *Quarterly Reviews of Biophysics* 36:307-340.

Williams, Bryony A. P. 2009. "Unique physiology of host–parasite interactions in microsporidia infections." *Cellular Microbiology* 11:1551-1560. doi: 10.1111/j.1462-5822.2009.01362.x.

Williams, Bryony A. P., and Patrick J. Keeling. 2011. "Microsporidia – Highly Reduced and Derived Relatives of Fungi." In *Evolution of Fungi and Fungal-Like Organisms*, 25-36. Springer, Berlin, Heidelberg.

Williams, Simon G., and Simon C. Lovell. 2009. "The Effect of Sequence Evolution on Protein Structural Divergence." *Molecular Biology and Evolution* 26:1055-1065. doi: 10.1093/molbev/msp020.

Wilson, Cyrus A., Julia Kreychman, and Mark Gerstein. 2000. "Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores." *Journal of Molecular Biology* 297:233-249. doi: 10.1006/jmbi.2000.3550.

Winkler, Herbert H., and H. Ekkehard Neuhaus. 1999. "Non-mitochondrial ATP transport." *Trends in Biochemical Sciences* 24:64-68. doi: 10.1016/S0968-0004(98)01334-6.

Wiredu Boakye, Dominic, Pattana Jaroenlak, Anuphap Prachumwat, Tom A. Williams, Kelly S. Bateman, Ornchuma Itsathitphaisarn, Kallaya Sritunyalucksana, Konrad H. Paszkiewicz, Karen A. Moore, Grant D. Stentiford, and Bryony A.P. Williams. 2017. "Decay of the glycolytic pathway and adaptation to intranuclear parasitism within Enterocytozoonidae microsporidia." *Environmental Microbiology* 19:2077-2089. doi: 10.1111/1462-2920.13734.

Woese, C R, O Kandler, and M L Wheelis. 1990. "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." *Proceedings of the National Academy of Sciences of the United States of America* 87:4576-4579.

Yin, Hang, ShaoPeng Wang, Yu-Hang Zhang, Yu-Dong Cai, and Hailin Liu. 2016. "Analysis of Important Gene Ontology Terms and Biological Pathways Related to Pancreatic Cancer." *BioMed Research International* 2016. doi: 10.1155/2016/7861274.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. 2008. "High-Quality Binary Protein Interaction Map of the Yeast Interactome Network." *Science* 322:104-110. doi: 10.1126/science.1158684.

Zhang, Chengxin, Peter L. Freddolino, and Yang Zhang. 2017. "COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information." *Nucleic Acids Research* 45:W291-W299. doi: 10.1093/nar/gkx366.

Zudilova-Seinstra, Elena, Tony Adriaansen, and Robert van Liere. 2009. "Overview of Interactive Visualization." In *Advanced Information and Knowledge Processing*, 3-15.

Zwickl, Derrick J., and David M. Hillis. 2002. "Increased Taxon Sampling Greatly Reduces Phylogenetic Error." *Systematic Biology* 51:588-598. doi: 10.1080/10635150290102339.

A. Appendix

Tables

Table A-1: Taxon set D - List of 491 species we used for the distribution analysis of microsporidian LCA proteins.

No.	Full name	Supertaxa	Group
1	*Ashbya gossypii*	Saccharomycotina	Fungi
2	*Candida albicans*	Saccharomycotina	Fungi
3	*Candida dubliniensis CD36*	Saccharomycotina	Fungi
4	*Candida glabrata*	Saccharomycotina	Fungi
5	*Candida parapsilosis*	Saccharomycotina	Fungi
6	*Candida tropicalis*	Saccharomycotina	Fungi
7	*Clavispora lusitaniae*	Saccharomycotina	Fungi
8	*Debaryomyces hansenii CBS767*	Saccharomycotina	Fungi
9	*Kluyveromyces lactis*	Saccharomycotina	Fungi
10	*Kluyveromyces thermotolerans*	Saccharomycotina	Fungi
11	*Kluyveromyces waltii*	Saccharomycotina	Fungi
12	*Lodderomyces elongisporus NRRL YB-4239*	Saccharomycotina	Fungi
13	*Pichia guilliermondii*	Saccharomycotina	Fungi
14	*Pichia pastoris GS115*	Saccharomycotina	Fungi
15	*Pichia stipitis CBS 6054*	Saccharomycotina	Fungi
16	*Saccharomyces bayanus 623-6C*	Saccharomycotina	Fungi
17	*Saccharomyces castelli*	Saccharomycotina	Fungi
18	*Saccharomyces cerevisiae*	Saccharomycotina	Fungi
19	*Saccharomyces kluyveri*	Saccharomycotina	Fungi
20	*Saccharomyces kudriavzevii*	Saccharomycotina	Fungi
21	*Saccharomyces mikatae*	Saccharomycotina	Fungi
22	*Saccharomyces paradoxus*	Saccharomycotina	Fungi
23	*Vanderwaltozyma polyspora*	Saccharomycotina	Fungi
24	*Yarrowia lipolytica*	Saccharomycotina	Fungi
25	*Zygosaccharomyces rouxii*	Saccharomycotina	Fungi
26	*Acidomyces richmondensis*	Pezizomycotina	Fungi
27	*Aulographum hederae*	Pezizomycotina	Fungi

28	*Baudoinia compniacensis uamh 10762*	Pezizomycotina	Fungi
29	*Botryosphaeria dothidea*	Pezizomycotina	Fungi
30	*Cenococcum geophilum 1.58*	Pezizomycotina	Fungi
31	*Cladonia grayi*	Pezizomycotina	Fungi
32	*Cochliobolus carbonum 26-r-13*	Pezizomycotina	Fungi
33	*Cochliobolus heterostrophus c5 3332*	Pezizomycotina	Fungi
34	*Cochliobolus heterostrophus c5 5759*	Pezizomycotina	Fungi
35	*Cochliobolus lunatus m118*	Pezizomycotina	Fungi
36	*Cochliobolus miyabeanus atcc 44560*	Pezizomycotina	Fungi
37	*Cochliobolus victoriae fi3*	Pezizomycotina	Fungi
38	*Cucurbitaria berberidis cbs 394.84*	Pezizomycotina	Fungi
39	*Dissoconium aciculare*	Pezizomycotina	Fungi
40	*Dothistroma septosporum nze10*	Pezizomycotina	Fungi
41	*Dothidotthia symphoricarpi*	Pezizomycotina	Fungi
42	*Hysterium pulicare*	Pezizomycotina	Fungi
43	*Leptosphaeria maculans*	Pezizomycotina	Fungi
44	*Lepidopterella palustris*	Pezizomycotina	Fungi
45	*Lophiostoma macrostomum*	Pezizomycotina	Fungi
46	*Macrophomina phaseolina ms6*	Pezizomycotina	Fungi
47	*Melanomma pulvis-pyrius*	Pezizomycotina	Fungi
48	*Myriangium duriaei cbs 260.36*	Pezizomycotina	Fungi
49	*Neofusicoccum parvum ucrnp2*	Pezizomycotina	Fungi
50	*Piedraia hortae*	Pezizomycotina	Fungi
51	*Pleomassaria siparia*	Pezizomycotina	Fungi
52	*Pyrenophora teres f. teres*	Pezizomycotina	Fungi
53	*Pyrenophora tritici-repentis pt-1c-bfp 3136*	Pezizomycotina	Fungi
54	*Pyrenophora tritici-repentis pt-1c-bfp 5809*	Pezizomycotina	Fungi
55	*Rhytidhysteron rufulum*	Pezizomycotina	Fungi
56	*Septoria musiva so2202*	Pezizomycotina	Fungi
57	*Septoria populicola*	Pezizomycotina	Fungi
58	*Thermomyces stellatus cbs 241.64*	Pezizomycotina	Fungi
59	*Trypethelium eluteriae*	Pezizomycotina	Fungi
60	*Zasmidium cellare atcc 36951*	Pezizomycotina	Fungi
61	*Zopfia rhizophila*	Pezizomycotina	Fungi

62	*Cladosporium fulvum*	Pezizomycotina	Fungi
63	*Cochliobolus sativus nd90pr*	Pezizomycotina	Fungi
64	*Didymella exigua cbs 183.55*	Pezizomycotina	Fungi
65	*Lentithecium fluviatile*	Pezizomycotina	Fungi
66	*Patellaria atrata*	Pezizomycotina	Fungi
67	*Polychaeton citri*	Pezizomycotina	Fungi
68	*Setosphaeria turcica et28a*	Pezizomycotina	Fungi
69	*Sporormia fimetaria*	Pezizomycotina	Fungi
70	*Xanthoria parietina*	Pezizomycotina	Fungi
71	*Ajellomyces capsulatus NAmI WU24*	Pezizomycotina	Fungi
72	*Ajellomyces dermatitidis ER-3*	Pezizomycotina	Fungi
73	*Alternaria brassicicola*	Pezizomycotina	Fungi
74	*Ascosphaera apis*	Pezizomycotina	Fungi
75	*Aspergillus clavatus*	Pezizomycotina	Fungi
76	*Aspergillus fischeri*	Pezizomycotina	Fungi
77	*Aspergillus flavus*	Pezizomycotina	Fungi
78	*Aspergillus fumigatus*	Pezizomycotina	Fungi
79	*Aspergillus kawachii*	Pezizomycotina	Fungi
80	*Aspergillus nidulans 2095*	Pezizomycotina	Fungi
81	*Aspergillus nidulans 1855*	Pezizomycotina	Fungi
82	*Aspergillus oryzae*	Pezizomycotina	Fungi
83	*Aspergillus terreus*	Pezizomycotina	Fungi
84	*Botrytis cinerea*	Pezizomycotina	Fungi
85	*Chaetomium globosum*	Pezizomycotina	Fungi
86	*Coccidioides immitis RS*	Pezizomycotina	Fungi
87	*Coccidioides posadasii RMSCC_3488*	Pezizomycotina	Fungi
88	*Cryphonectria parasitica 3352*	Pezizomycotina	Fungi
89	*Cryphonectria parasitica 4119*	Pezizomycotina	Fungi
90	*Fusarium graminearum ph1*	Pezizomycotina	Fungi
91	*Fusarium oxysporum lycopersici*	Pezizomycotina	Fungi
92	*Fusarium verticillioides*	Pezizomycotina	Fungi
93	*Magnaporthe grisea*	Pezizomycotina	Fungi
94	*Microsporum canis CBS 113480*	Pezizomycotina	Fungi
95	*Microsporum gypseum CBS 118893*	Pezizomycotina	Fungi

96	*Mycosphaerella fijiensis*	Pezizomycotina	Fungi
97	*Mycosphaerella graminicola*	Pezizomycotina	Fungi
98	*Nectria haematococca MPVI*	Pezizomycotina	Fungi
99	*Neurospora crassa*	Pezizomycotina	Fungi
100	*Neurospora discreta FGSC 8579 mat A*	Pezizomycotina	Fungi
101	*Neurospora tetrasperma FGSC 2508 mat A*	Pezizomycotina	Fungi
102	*Paracoccidioides brasiliensis Pb03*	Pezizomycotina	Fungi
103	*Penicillium chrysogenum*	Pezizomycotina	Fungi
104	*Penicillium marneffei ATCC 18224*	Pezizomycotina	Fungi
105	*Podospora anserina*	Pezizomycotina	Fungi
106	*Sclerotinia sclerotiorum*	Pezizomycotina	Fungi
107	*Stagonospora nodorum*	Pezizomycotina	Fungi
108	*Talaromyces stipitatus*	Pezizomycotina	Fungi
109	*Thielavia terrestris*	Pezizomycotina	Fungi
110	*Trichoderma atroviride*	Pezizomycotina	Fungi
111	*Trichophyton equinum CBS127.97*	Pezizomycotina	Fungi
112	*Trichoderma reesei*	Pezizomycotina	Fungi
113	*Trichoderma virens Gv29-8*	Pezizomycotina	Fungi
114	*Tuber melanosporum*	Pezizomycotina	Fungi
115	*Uncinocarpus reesii 5820*	Pezizomycotina	Fungi
116	*Uncinocarpus reesii 2939*	Pezizomycotina	Fungi
117	*Verticillium albo-atrum VaMs.102*	Pezizomycotina	Fungi
118	*Verticillium dahliae VdLs.17*	Pezizomycotina	Fungi
119	*Phaeosphaeria nodorum SN15*	Pezizomycotina	Fungi
120	*Schizosaccharomyces japonicus*	Taphrinomycotina	Fungi
121	*Schizosaccharomyce octosporus*	Taphrinomycotina	Fungi
122	*Schizosaccharomyces pombe*	Taphrinomycotina	Fungi
123	*Schizosaccharomyces sp. OY26*	Taphrinomycotina	Fungi
124	*Coprinopsis cinerea*	Basidiomycota	Fungi
125	*Cryptococcus neoformans JEC21*	Basidiomycota	Fungi
126	*Gelatoporia subvermispora*	Basidiomycota	Fungi
127	*Heterobasidion annosum*	Basidiomycota	Fungi
128	*Laccaria bicolor*	Basidiomycota	Fungi
129	*Malassezia globosa CBS 7966*	Basidiomycota	Fungi

130	*Melampsora laricis-populina*	Basidiomycota	Fungi
131	*Moniliophthora perniciosa FA553*	Basidiomycota	Fungi
132	*Phanerochaete chrysosporium P-78*	Basidiomycota	Fungi
133	*Pleurotus ostreatus PC15*	Basidiomycota	Fungi
134	*Postia placenta*	Basidiomycota	Fungi
135	*Puccinia graminis*	Basidiomycota	Fungi
136	*Schizophyllum commune*	Basidiomycota	Fungi
137	*Serpula lacrymans S7_3*	Basidiomycota	Fungi
138	*Sporobolomyces roseus*	Basidiomycota	Fungi
139	*Tremella mesenterica Fries*	Basidiomycota	Fungi
140	*Ustilago maydis*	Basidiomycota	Fungi
141	*Mucor circinelloides*	Mucoromycotina	Fungi
142	*Phycomyces blakesleeanus*	Mucoromycotina	Fungi
143	*Rhizopus oryzae*	Mucoromycotina	Fungi
144	*Batrachochytrium dendrobatidis*	Chytridiomycota	Fungi
145	*Spizellomyces punctatus*	Chytridiomycota	Fungi
146	*Encephalitozoon hellem*	Microsporidia	Microsporidia
147	*Encephalitozoon intestinalis*	Microsporidia	Microsporidia
148	*Encephalitozoon cuniculi*	Microsporidia	Microsporidia
149	*Nosema ceranae*	Microsporidia	Microsporidia
150	*Enterocytozoon bieneusi*	Microsporidia	Microsporidia
151	*Antonospora locustae*	Microsporidia	Microsporidia
152	*Edhazardia aedis*	Microsporidia	Microsporidia
153	*Vavraia culicis floridensis*	Microsporidia	Microsporidia
154	*Nematocida parisii*	Microsporidia	Microsporidia
155	*Anncaliia algerae PRA339*	Microsporidia	Microsporidia
156	*Vittaforma corneae*	Microsporidia	Microsporidia
157	*Anas platyrhynchos*	Metazoa	Unikonta
158	*Latimeria chalumnae*	Metazoa	Unikonta
159	*mustela putorius furo*	Metazoa	Unikonta
160	*Linepithema humile*	Metazoa	Unikonta
161	*Pelodiscus sinensis*	Metazoa	Unikonta
162	*Acropora digitifera*	Metazoa	Unikonta
163	*Acyrthosiphon pisum*	Metazoa	Unikonta

164	*Aedes aegypti*	Metazoa	Unikonta
165	*Ailuropoda melanoleuca*	Metazoa	Unikonta
166	*Amphimedon queenslandica*	Metazoa	Unikonta
167	*Anolis carolinensis*	Metazoa	Unikonta
168	*Anopheles gambiae*	Metazoa	Unikonta
169	*Apis mellifera*	Metazoa	Unikonta
170	*Bombyx mori*	Metazoa	Unikonta
171	*Bos taurus*	Metazoa	Unikonta
172	*Branchiostoma floridae*	Metazoa	Unikonta
173	*Caenorhabditis brenneri 2851*	Metazoa	Unikonta
174	*Caenorhabditis brenneri 70*	Metazoa	Unikonta
175	*Caenorhabditis elegans*	Metazoa	Unikonta
176	*Caenorhabditis japonica*	Metazoa	Unikonta
177	*Caenorhabditis remanei*	Metazoa	Unikonta
178	*Callithrix jacchus*	Metazoa	Unikonta
179	*Canis familiaris*	Metazoa	Unikonta
180	*Capitella capitata*	Metazoa	Unikonta
181	*Cavia porcellus*	Metazoa	Unikonta
182	*Choloepus hoffmanni*	Metazoa	Unikonta
183	*Ciona intestinalis*	Metazoa	Unikonta
184	*Ciona savignyi*	Metazoa	Unikonta
185	*Culex pipiens quinquefasciatus*	Metazoa	Unikonta
186	*Danio rerio*	Metazoa	Unikonta
187	*Daphnia pulex*	Metazoa	Unikonta
188	*Dasypus novemcinctus*	Metazoa	Unikonta
189	*Dipodomys ordii*	Metazoa	Unikonta
190	*Drosophila ananassae*	Metazoa	Unikonta
191	*Drosophila erecta*	Metazoa	Unikonta
192	*Drosophila grimshawi*	Metazoa	Unikonta
193	*Drosophila melanogaster*	Metazoa	Unikonta
194	*Drosophila mojavensis*	Metazoa	Unikonta
195	*Drosophila persimilis*	Metazoa	Unikonta
196	*Drosophila pseudoobscura*	Metazoa	Unikonta
197	*Drosophila sechellia*	Metazoa	Unikonta

198	*Drosophila simulans*	Metazoa	Unikonta
199	*Drosophila virilis*	Metazoa	Unikonta
200	*Drosophila willistoni*	Metazoa	Unikonta
201	*Drosophila yakuba*	Metazoa	Unikonta
202	*Echinops telfairi*	Metazoa	Unikonta
203	*Equus caballus*	Metazoa	Unikonta
204	*Erinaceus europaeus*	Metazoa	Unikonta
205	*Felis catus*	Metazoa	Unikonta
206	*Takifugu rubripes*	Metazoa	Unikonta
207	*Gadus morhua*	Metazoa	Unikonta
208	*Gallus gallus*	Metazoa	Unikonta
209	*Gasterosteus aculeatus*	Metazoa	Unikonta
210	*Gorilla gorilla*	Metazoa	Unikonta
211	*Helobdella robusta*	Metazoa	Unikonta
212	*Homo sapiens*	Metazoa	Unikonta
213	*Hydra magnipapillata*	Metazoa	Unikonta
214	*Ixodes scapularis*	Metazoa	Unikonta
215	*Lama pacos*	Metazoa	Unikonta
216	*Lepisosteus oculatus*	Metazoa	Unikonta
217	*Loa loa*	Metazoa	Unikonta
218	*Lottia gigantea*	Metazoa	Unikonta
219	*Loxodonta africana*	Metazoa	Unikonta
220	*Macropus eugenii*	Metazoa	Unikonta
221	*Macaca mulatta*	Metazoa	Unikonta
222	*Microcebus murinus*	Metazoa	Unikonta
223	*Monodelphis domestica*	Metazoa	Unikonta
224	*Mus musculus*	Metazoa	Unikonta
225	*Myotis lucifugus*	Metazoa	Unikonta
226	*Nasonia vitripennis*	Metazoa	Unikonta
227	*Nematostella vectensis*	Metazoa	Unikonta
228	*Nomascus leucogenys*	Metazoa	Unikonta
229	*Ochotona princeps*	Metazoa	Unikonta
230	*Ornithorhynchus anatinus*	Metazoa	Unikonta
231	*Oryctolagus cuniculus*	Metazoa	Unikonta

232	*Oryzias latipes*	Metazoa	Unikonta
233	*Otolemur garnettii*	Metazoa	Unikonta
234	*Pan troglodytes*	Metazoa	Unikonta
235	*Pediculus humanus*	Metazoa	Unikonta
236	*Petromyzon marinus*	Metazoa	Unikonta
237	*Pongo pygmaeus*	Metazoa	Unikonta
238	*Pristionchus pacificus*	Metazoa	Unikonta
239	*Procavia capensis*	Metazoa	Unikonta
240	*Pteropus vampyrus*	Metazoa	Unikonta
241	*Rattus norvegicus*	Metazoa	Unikonta
242	*Sarcophilus_harrisii*	Metazoa	Unikonta
243	*Schistosoma mansoni*	Metazoa	Unikonta
244	*Sorex araneus*	Metazoa	Unikonta
245	*Spermophilus tridecemlineatus*	Metazoa	Unikonta
246	*Strongylocentrotus purpuratus*	Metazoa	Unikonta
247	*Sus scrofa*	Metazoa	Unikonta
248	*Taeniopygia guttata*	Metazoa	Unikonta
249	*Tarsius syrichta*	Metazoa	Unikonta
250	*Tetraodon nigroviridis*	Metazoa	Unikonta
251	*Trichoplax adhaerens*	Metazoa	Unikonta
252	*Tribolium castaneum*	Metazoa	Unikonta
253	*Tupaia belangeri*	Metazoa	Unikonta
254	*Tursiops truncatus*	Metazoa	Unikonta
255	*Wuchereria bancrofti*	Metazoa	Unikonta
256	*Xenopus tropicalis*	Metazoa	Unikonta
257	*Callorhinchus milii*	Metazoa	Unikonta
258	*Monosiga brevicollis*	-	Unikonta
259	*Capsaspora owczarzaki*	-	Unikonta
260	*Thecamonas trahens*	Thecamonas_trahens	Unikonta
261	*Bigelowiella natans*	Amoebozoa	Unikonta
262	*Dictyostelium discoideum AX4*	Amoebozoa	Unikonta
263	*Dictyostelium purpureum QSDP1*	Amoebozoa	Unikonta
264	*Entamoeba dispar SAW760*	Amoebozoa	Unikonta
265	*Entamoeba histolytica*	Amoebozoa	Unikonta

266	*Polysphondylium pallidum*	Amoebozoa	Unikonta
267	*Leishmania braziliensis*	Euglenozoa	Eukaryota
268	*Leishmania infantum*	Euglenozoa	Eukaryota
269	*Leishmania major strain Friedlin*	Euglenozoa	Eukaryota
270	*Trypanosoma brucei*	Euglenozoa	Eukaryota
271	*Naegleria gruberi*	Heterolobosea	Eukaryota
272	*Aquilegia coerulea*	Streptophyta	Eukaryota
273	*Arabidopsis lyrata*	Streptophyta	Eukaryota
274	*Arabidopsis thaliana*	Streptophyta	Eukaryota
275	*Brachypodium distachyon*	Streptophyta	Eukaryota
276	*Brassica rapa*	Streptophyta	Eukaryota
277	*Capsella rubella*	Streptophyta	Eukaryota
278	*Citrus clementina*	Streptophyta	Eukaryota
279	*Citrus sinensis*	Streptophyta	Eukaryota
280	*Cucumis sativus*	Streptophyta	Eukaryota
281	*Eucalyptus grandis*	Streptophyta	Eukaryota
282	*Glycine max*	Streptophyta	Eukaryota
283	*Linum usitatissimum*	Streptophyta	Eukaryota
284	*Malus x domestica*	Streptophyta	Eukaryota
285	*Manihot esculenta*	Streptophyta	Eukaryota
286	*Medicago truncatula*	Streptophyta	Eukaryota
287	*Mimulus guttatus*	Streptophyta	Eukaryota
288	*Oryza sativa sp. japonica*	Streptophyta	Eukaryota
289	*Phaseolus vulgaris*	Streptophyta	Eukaryota
290	*Physcomitrella patens sp. patens*	Streptophyta	Eukaryota
291	*Populus trichocarpa*	Streptophyta	Eukaryota
292	*Prunus persica*	Streptophyta	Eukaryota
293	*Ricinus communis*	Streptophyta	Eukaryota
294	*Selaginella moellendorffii*	Streptophyta	Eukaryota
295	*Setaria italica*	Streptophyta	Eukaryota
296	*Solanum lycopersicum*	Streptophyta	Eukaryota
297	*Sorghum bicolor*	Streptophyta	Eukaryota
298	*Vitis vinifera*	Streptophyta	Eukaryota
299	*Zea mays*	Streptophyta	Eukaryota

300	*Thellungiella halophila*	Streptophyta	Eukaryota
301	*Chlorella sp. NC64A*	Chlorophyta	Eukaryota
302	*Chlamydomonas reinhardtii*	Chlorophyta	Eukaryota
303	*Micromonas sp. CCMP490*	Chlorophyta	Eukaryota
304	*Micromonas pusilla sp. RCC299*	Chlorophyta	Eukaryota
305	*Ostreococcus lucimarinus*	Chlorophyta	Eukaryota
306	*Ostreococcus sp. RCC809*	Chlorophyta	Eukaryota
307	*Ostreococcus tauri*	Chlorophyta	Eukaryota
308	*Volvox carteri f. nagariensis*	Chlorophyta	Eukaryota
309	*Coccomyxa subellipsoidea*	Chlorophyta	Eukaryota
310	*Cyanidioschyzon merolae*	Rhodophyta	Eukaryota
311	*Aureococcus anophagefferens*	Stramenopiles	Eukaryota
312	*Ectocarpus siliculosus*	Stramenopiles	Eukaryota
313	*Fragilariopsis cylindrus CCMP 1102*	Stramenopiles	Eukaryota
314	*Phaeodactylum tricornutum*	Stramenopiles	Eukaryota
315	*Phytophthora infestans*	Stramenopiles	Eukaryota
316	*Phytophthora ramorum*	Stramenopiles	Eukaryota
317	*Phytophthora sojae*	Stramenopiles	Eukaryota
318	*Saprolegnia parasitica*	Stramenopiles	Eukaryota
319	*Thalassiosira pseudonana*	Stramenopiles	Eukaryota
320	*Babesia bovis*	Alveolata	Eukaryota
321	*Cryptosporidium hominis ATCC BAA-381*	Alveolata	Eukaryota
322	*Eimeria tenella*	Alveolata	Eukaryota
323	*Neospora caninum*	Alveolata	Eukaryota
324	*Paramecium tetraurelia*	Alveolata	Eukaryota
325	*Perkinsus marinus*	Alveolata	Eukaryota
326	*Plasmodium berghei*	Alveolata	Eukaryota
327	*Plasmodium chabaudi*	Alveolata	Eukaryota
328	*Plasmodium falciparum*	Alveolata	Eukaryota
329	*Plasmodium gallinaceum*	Alveolata	Eukaryota
330	*Plasmodium knowlesi*	Alveolata	Eukaryota
331	*Plasmodium reichenowi*	Alveolata	Eukaryota
332	*Plasmodium vivax*	Alveolata	Eukaryota
333	*Plasmodium yoelii*	Alveolata	Eukaryota

334	*Tetrahymena thermophila*	Alveolata	Eukaryota
335	*Theileria annulata*	Alveolata	Eukaryota
336	*Theileria parva*	Alveolata	Eukaryota
337	*Toxoplasma gondii*	Alveolata	Eukaryota
338	*Emiliania huxleyi CCMP1516*	Haptophyceae	Eukaryota
339	*Hemiselmis andersenii*	Cryptophyta	Eukaryota
340	*Guillardia theta*	Cryptophyta	Eukaryota
341	*Hemiselmis andersenii*	Cryptophyta	Eukaryota
342	*Archaeoglobus fulgidus*	Euryarchaeota	Archaea
343	*Methanococcoides burtonii*	Euryarchaeota	Archaea
344	*Methanopyrus kandleri*	Euryarchaeota	Archaea
345	*Methanocorpusculum labreanum*	Euryarchaeota	Archaea
346	*Natronomonas pharaonis*	Euryarchaeota	Archaea
347	*Haloferax volcanii DS2*	Euryarchaeota	Archaea
348	*Methanosarcina barkeri str. Fusaro*	Euryarchaeota	Archaea
349	*Methanocaldococcus jannaschii DSM 2661*	Euryarchaeota	Archaea
350	*Methanothermobacter thermautotrophicus str. Delta H*	Euryarchaeota	Archaea
351	*Picrophilus torridus DSM 9790*	Euryarchaeota	Archaea
352	*Pyrococcus horikoshii*	Euryarchaeota	Archaea
353	*Thermoplasma acidophilum DSM 1728*	Euryarchaeota	Archaea
354	*Thermococcus kodakarensis KOD1*	Euryarchaeota	Archaea
355	*Nanoarchaeum equitans*	Nanoarchaeota	Archaea
356	*Candidatus Korarchaeum cryptofilum OPF8*	Korarchaeota	Archaea
357	*Aeropyrum pernix K1*	Crenarchaeota	Archaea
358	*Ignicoccus hospitalis*	Crenarchaeota	Archaea
359	*Metallosphaera sedula*	Crenarchaeota	Archaea
360	*Pyrobaculum neutrophilum*	Crenarchaeota	Archaea
361	*Thermofilum pendens*	Crenarchaeota	Archaea
362	*Caldivirga maquilingensis*	Crenarchaeota	Archaea
363	*Sulfolobus solfataricus P2*	Crenarchaeota	Archaea
364	*Candidatus Caldiarchaeum subterraneum*	Thaumarchaeota	Archaea
365	*Cenarchaeum symbiosum*	Thaumarchaeota	Archaea
366	*Nitrosopumilus maritimus*	Thaumarchaeota	Archaea

367	*Candidatus Nitrososphaera gargensis Ga9.2*	Thaumarchaeota	Archaea
368	*Deinococcus proteolyticus MRP*	*Deinococci*	Bacteria
369	*Marinithermus hydrothermalis DSM 14884*	*Deinococci*	Bacteria
370	*Clostridium tetani E88*	*Firmicutes*	Bacteria
371	*Coprothermobacter proteolyticus DSM 5265*	*Firmicutes*	Bacteria
372	*Desulfotomaculum acetoxidans DSM 771*	*Firmicutes*	Bacteria
373	*Acaryochloris marina*	*Cyanobacteria*	Bacteria
374	*Acaryochloris marina*	*Cyanobacteria*	Bacteria
375	*Anabaena cylindrica*	*Cyanobacteria*	Bacteria
376	*Anabaena sp.*	*Cyanobacteria*	Bacteria
377	*Anabaena variabilis ATCC 29413*	*Cyanobacteria*	Bacteria
378	*Arthrospira platensis*	*Cyanobacteria*	Bacteria
379	*Calothrix sp. 5685*	*Cyanobacteria*	Bacteria
380	*Calothrix sp. 5686*	*Cyanobacteria*	Bacteria
381	*Chamaesiphon minutus*	*Cyanobacteria*	Bacteria
382	*Chlorogloeopsis fritschii*	*Cyanobacteria*	Bacteria
383	*Chlorogloeopsis sp.*	*Cyanobacteria*	Bacteria
384	*Chroococcidiopsis thermalis*	*Cyanobacteria*	Bacteria
385	*Crinalium epipsammum*	*Cyanobacteria*	Bacteria
386	*Cyanobacterium aponinum*	*Cyanobacteria*	Bacteria
387	*Cyanothece ATCC 51142*	*Cyanobacteria*	Bacteria
388	*Cyanobium gracile*	*Cyanobacteria*	Bacteria
389	*Cyanothece sp. 5693*	*Cyanobacteria*	Bacteria
390	*Cyanothece sp. 5694*	*Cyanobacteria*	Bacteria
391	*Cyanothece sp. 5695*	*Cyanobacteria*	Bacteria
392	*Cyanothece sp. 5696*	*Cyanobacteria*	Bacteria
393	*Cyanothece sp. 5697*	*Cyanobacteria*	Bacteria
394	*Cyanothece sp. 5698*	*Cyanobacteria*	Bacteria
395	*Cyanobacterium stanieri*	*Cyanobacteria*	Bacteria
396	*Cyanobacterium UCYN-A*	*Cyanobacteria*	Bacteria
397	*Cylindrospermum stagnale*	*Cyanobacteria*	Bacteria
398	*Dactylococcopsis salina*	*Cyanobacteria*	Bacteria
399	*Fischerella muscicola 5744*	*Cyanobacteria*	Bacteria
400	*Fischerella muscicola 5745*	*Cyanobacteria*	Bacteria

401	*Fischerella sp.*	*Cyanobacteria*	Bacteria
402	*Geitlerinema sp.*	*Cyanobacteria*	Bacteria
403	*Gloeocapsa sp.*	*Cyanobacteria*	Bacteria
404	*Gloeobacter violaceus 4698*	*Cyanobacteria*	Bacteria
405	*Gloeobacter violaceus 5702*	*Cyanobacteria*	Bacteria
406	*Halothece sp.*	*Cyanobacteria*	Bacteria
407	*Leptolyngbya sp.*	*Cyanobacteria*	Bacteria
408	*Microcystis aeruginosa NIES 843*	*Cyanobacteria*	Bacteria
409	*Microcoleus sp.*	*Cyanobacteria*	Bacteria
410	*Nostoc azollae 0708*	*Cyanobacteria*	Bacteria
411	*Nostoc punctiforme PCC 73102*	*Cyanobacteria*	Bacteria
412	*Nostoc sp. 5707*	*Cyanobacteria*	Bacteria
413	*Nostoc sp. 5708*	*Cyanobacteria*	Bacteria
414	*Nostoc sp. 5709*	*Cyanobacteria*	Bacteria
415	*Oscillatoria acuminata*	*Cyanobacteria*	Bacteria
416	*Oscillatoria nigro-viridis*	*Cyanobacteria*	Bacteria
417	*Pleurocapsa sp.*	*Cyanobacteria*	Bacteria
418	*Prochlorococcus marinus AS9601 4702*	*Cyanobacteria*	Bacteria
419	*Prochlorococcus marinus AS9601 5713*	*Cyanobacteria*	Bacteria
420	*Prochlorococcus marinus AS9601 5714*	*Cyanobacteria*	Bacteria
421	*Prochlorococcus marinus AS9601 5715*	*Cyanobacteria*	Bacteria
422	*Prochlorococcus marinus AS9601 5716*	*Cyanobacteria*	Bacteria
423	*Prochlorococcus marinus AS9601 5717*	*Cyanobacteria*	Bacteria
424	*Prochlorococcus marinus AS9601 5718*	*Cyanobacteria*	Bacteria
425	*Prochlorococcus marinus AS9601 5719*	*Cyanobacteria*	Bacteria
426	*Prochlorococcus marinus AS9601 5720*	*Cyanobacteria*	Bacteria
427	*Prochlorococcus marinus AS9601 5721*	*Cyanobacteria*	Bacteria
428	*Prochlorococcus marinus AS9601 5722*	*Cyanobacteria*	Bacteria
429	*Prochlorococcus marinus AS9601 5723*	*Cyanobacteria*	Bacteria
430	*Prochlorococcus marinus AS9601 5724*	*Cyanobacteria*	Bacteria
431	*Pseudanabaena sp.*	*Cyanobacteria*	Bacteria
432	*Rivularia sp.*	*Cyanobacteria*	Bacteria
433	*Scytonema hofmanni*	*Cyanobacteria*	Bacteria
434	*Stanieria cyanosphaera*	*Cyanobacteria*	Bacteria

435	*Synechococcus elongatus PCC 7942 4703*	*Cyanobacteria*	Bacteria
436	*Synechococcus elongatus PCC 7942 4704*	*Cyanobacteria*	Bacteria
437	*Synechococcus_sp_JA-2-3Ba_2-13 4694*	*Cyanobacteria*	Bacteria
438	*Synechococcus_sp_JA-2-3Ba_2-13 4695*	*Cyanobacteria*	Bacteria
439	*Synechocystis sp. 5728*	*Cyanobacteria*	Bacteria
440	*Synechocystis sp. 5729*	*Cyanobacteria*	Bacteria
441	*Synechocystis sp. 5730*	*Cyanobacteria*	Bacteria
442	*Synechocystis sp. 5731*	*Cyanobacteria*	Bacteria
443	*Synechocystis sp. 5731*	*Cyanobacteria*	Bacteria
444	*Synechocystis sp. 5733*	*Cyanobacteria*	Bacteria
445	*Synechocystis sp. 5734*	*Cyanobacteria*	Bacteria
446	*Synechocystis sp. 5735*	*Cyanobacteria*	Bacteria
447	*Synechocystis sp. 5736*	*Cyanobacteria*	Bacteria
448	*Synechocystis sp. 5737*	*Cyanobacteria*	Bacteria
449	*Synechocystis sp. 5738*	*Cyanobacteria*	Bacteria
450	*Synechocystis sp. 5739*	*Cyanobacteria*	Bacteria
451	*Synechocystis sp. 5740*	*Cyanobacteria*	Bacteria
452	*Thermosynechococcus elongatus 4705*	*Cyanobacteria*	Bacteria
453	*Thermosynechococcus elongatus 5741*	*Cyanobacteria*	Bacteria
454	*Trichodesmium erythraeum IMS101*	*Cyanobacteria*	Bacteria
455	*Clavibacter michiganensis subsp. michiganensis NCPPB 382*	*Actinobacteria*	Bacteria
456	*Conexibacter woesei DSM 14684*	*Actinobacteria*	Bacteria
457	*Chlamydophila psittaci 6BC*	*Chlamydiae*	Bacteria
458	*Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2*	*Bacteroidetes*	Bacteria
459	*Candidatus Sulcia muelleri DMIN*	*Bacteroidetes*	Bacteria
460	*Campylobacter curvus 525.92*	*Epsilonproteobacteria*	Bacteria
461	*Nitratiruptor sp. SB155-2*	*Epsilonproteobacteria*	Bacteria
462	*Sulfurovum sp. NBC37-1*	*Epsilonproteobacteria*	Bacteria
463	*Bdellovibrio bacteriovorus HD100*	*Deltaproteobacteria*	Bacteria
464	*Desulfovibrio vulgaris DP4*	*Deltaproteobacteria*	Bacteria
465	*Geobacter sulfurreducens PCA*	*Deltaproteobacteria*	Bacteria
466	*Sorangium cellulosum So ce 56*	*Deltaproteobacteria*	Bacteria

467	*Syntrophus aciditrophicus SB*	*Deltaproteobacteria*	Bacteria
468	*Agrobacterium fabrum*	*Alphaproteobacteria*	Bacteria
469	*Caulobacter crescentus CB15*	*Alphaproteobacteria*	Bacteria
470	*Ehrlichia canis str. Jake*	*Alphaproteobacteria*	Bacteria
471	*Maricaulis maris MCS10*	*Alphaproteobacteria*	Bacteria
472	*Zymomonas mobilis subsp. mobilis ZM4*	*Alphaproteobacteria*	Bacteria
473	*Bordetella petrii DSM 12804*	*Betaproteobacteria*	Bacteria
474	*Chlamydia trachomatis G/9301*	*Betaproteobacteria*	Bacteria
475	*Dechloromonas aromatica RCB*	*Betaproteobacteria*	Bacteria
476	*Methylobacillus flagellatus KT*	*Betaproteobacteria*	Bacteria
477	*Neisseria gonorrhoeae FA 1090*	*Betaproteobacteria*	Bacteria
478	*Nitrosomonas europaea ATCC 19718*	*Betaproteobacteria*	Bacteria
479	*Thiobacillus denitrificans ATCC 25259*	*Betaproteobacteria*	Bacteria
480	*Aeromonas hydrophila subsp. hydrophila ATCC 7966*	*Gammaproteobacteria*	Bacteria
481	*Baumannia cicadellinicola str. Hc (Homalodisca coagulata)*	*Gammaproteobacteria*	Bacteria
482	*Candidatus Carsonella ruddii PV*	*Gammaproteobacteria*	Bacteria
483	*Coxiella burnetii RSA 331*	*Gammaproteobacteria*	Bacteria
484	*Dichelobacter nodosus VCS1703A*	*Gammaproteobacteria*	Bacteria
485	*Escherichia coli str. K-12 substr. MG1655*	*Gammaproteobacteria*	Bacteria
486	*Haemophilus influenzae 10810*	*Gammaproteobacteria*	Bacteria
487	*Marinomonas mediterranea MMB-1*	*Gammaproteobacteria*	Bacteria
488	*Methylococcus capsulatus str. Bath*	*Gammaproteobacteria*	Bacteria
489	*Nitrosococcus oceani ATCC 19707*	*Gammaproteobacteria*	Bacteria
490	*Pseudomonas putida F1*	*Gammaproteobacteria*	Bacteria
491	*Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)*	*Gammaproteobacteria*	Bacteria

Table A-2: Taxon set C - 72 taxa used for fungal tree reconstruction.

No.	ID	Taxon name	Genus	Order	Class	Phylum	Source
1	5037	*Ajellomyces capsulatus*	*Histoplasma*	Onygenales	Eurotiomycetes	Ascomycota	NCBI
2	746128	*Aspergillus fumigatus*	*Aspergillus*	Eurotiales	Eurotiomycetes	Ascomycota	NCBI
3	162425	*Aspergillus nidulans*	*Aspergillus*	Eurotiales	Eurotiomycetes	Ascomycota	NCBI
4	34373	*Blumeria graminis*	*Blumeria*	Erysiphales	Leotiomycetes	Ascomycota	NCBI
5	5476	*Candida albicans*	*Candida*	Saccharomycetales	Saccharomycetes	Ascomycota	NCBI
6	5501	*Coccidioides immitis*	*Coccidioides*	Onygenales	Eurotiomycetes	Ascomycota	NCBI
7	5507	*Fusarium oxysporum*	*Fusarium*	Hypocreales	Sordariomycetes	Ascomycota	NCBI
8	318829	*Magnaporthe oryzae*	*Magnaporthe*	Magnaporthales	Sordariomycetes	Ascomycota	NCBI
9	42068	*Pneumocystis jirovecii*	*Pneumocystis*	Pneumocystidales	Pneumocystidomycetes	Ascomycota	NCBI
10	4932	*Saccharomyces cerevisiae*	*Saccharomyces*	Saccharomycetales	Saccharomycetes	Ascomycota	NCBI
11	4896	*Schizosaccharomyces pombe*	*Schizosaccharomyces*	Schizosaccharomycetales	Schizosaccharomycetes	Ascomycota	NCBI
12	5551	*Trichophyton rubrum*	*Trichophyton*	Onygenales	Eurotiomycetes	Ascomycota	NCBI
13	5346	*Coprinopsis cinerea*	*Coprinopsis*	Agaricales	Agaricomycetes	Basidiomycota	NCBI
14	5207	*Cryptococcus neoformans*	*Filobasidiella*	Tremellales	Tremellomycetes	Basidiomycota	NCBI
15	76773	*Malassezia globosa*	*Malassezia*	Malasseziales	Malasseziomycetes	Basidiomycota	NCBI
16	203908	*Melampsora larici-*	*Melampsora*	Pucciniales	Pucciniomycetes	Basidiomycota	NCBI

		populina					
17	221103	*Moniliophthora roreri*	*Moniliophthora*	Agaricales	Agaricomycetes	Basidiomycota	NCBI
18	231932	*Phanerochaete carnosa*	*Phanerochaete*	Polyporales	Agaricomycetes	Basidiomycota	NCBI
19	5306	*Phanerochaete chrysosporium*	*Phanerochaete*	Polyporales	Agaricomycetes	Basidiomycota	NCBI
20	65672	*Piriformospora indica*	*Piriformospora*	Sebacinales	Agaricomycetes	Basidiomycota	NCBI
21	104341	*Postia placenta*	*Postia*	Polyporales	Agaricomycetes	Basidiomycota	NCBI
22	5297	*Puccinia graminis*	*Puccinia*	Pucciniales	Pucciniomycetes	Basidiomycota	NCBI
23	456999	*Rhizoctonia solani*	*Rhizoctonia*	Cantharellales	Agaricomycetes	Basidiomycota	NCBI
24	5334	*Schizophyllum commune*	*Schizophyllum*	Agaricales	Agaricomycetes	Basidiomycota	NCBI
25	85982	*Serpula lacrymans*	*Serpula*	Boletales	Agaricomycetes	Basidiomycota	NCBI
26	72558	*Sporisorium reilianum*	*Sporisorium*	Ustilaginales	Ustilaginomycetes	Basidiomycota	NCBI
27	40492	*Stereum hirsutum*	*Stereum*	Russulales	Agaricomycetes	Basidiomycota	NCBI
28	5217	*Tremella mesenterica*	*Tremella*	Tremellales	Tremellomycetes	Basidiomycota	NCBI
29	5270	*Ustilago maydis*	*Ustilago*	Ustilaginales	Ustilaginomycetes	Basidiomycota	NCBI
30	28583	*Allomyces macrogynus*	*Allomyces*	Blastocladiales	Blastocladiomycetes	Blastocladiomycota	Broad
31	109871	*Batrachochytrium dendrobatidis*	*Batrachochytrium*	Rhizophydiales	Chytridiomycetes	Chytridiomycota	JGI
32	166479	*Homoloaphlyctis polyrhiza JEL 142*	*Homolaphlyctis*	Rhizophydiales	Chytridiomycetes	Chytridiomycota	fungalgenomes.org

33	1123529	*Gonapodya prolifera*	*Gonapodya*	Monoblepharidales	Monoblepharidomycetes	Chytridiomycota	JGI
34	645134	*Spizellomyces punctatus DAOM BR117*	*Spizellomyces*	Spizellomycetales	Chytridiomycetes	Chytridiomycota	Broad
35	281847	*Rozella allomycis*	*Rozella*	NA	NA	Cryptomycota	NCBI
36	34488	*Conidiobolus coronatus*	*Conidiobolus*	Entomophthorales	Entomophthoromycetes	Entomophthoromycota	JGI
37	588596	*Rhizophagus irregularis*	*Rhizophagus*	Glomerales	Glomeromycetes	Glomeromycota	JGI
38	1004703	*Orpinomyces sp.*	*Orpinomyces*	Neocallimastigales	Neocallimastigomycetes	Neocallimastigomycota	JGI
39	45796	*Piromyces sp.*	*Piromyces*	Neocallimastigales	Neocallimastigomycetes	Neocallimastigomycota	JGI
40	61392	*Coemansia reversa*	*Coemansia*	Kickxellales	NA	NA	JGI
41	64518	*Mortierella alpina*	*Mortierella*	Mortierellales	NA	NA	phylomedb
42	78898	*Mortierella verticillata*	*Mortierella*	Mortierellales	NA	NA	JGI
43	36080	*Mucor circinelloides*	*Mucor*	Mucorales	NA	NA	JGI
44	4837	*Phycomyces blakesleeanus*	*Phycomyces*	Mucorales	NA	NA	JGI
45	64495	*Rhizopus oryzae RA 99-880*	*Rhizopus*	Mucorales	NA	NA	JGI
46	420593	*Lichtheimia hyalospora*	*Lichtheimia*	Mucorales	NA	NA	141.35.171.48
47	42458	*Lichtheimia corymbifera*	*Lichtheimia*	Mucorales	NA	NA	141.35.171.48

48	688394	*Lichtheimia ramosa*	*Lichtheimia*	Mucorales	NA	NA	141.35.171.48
49	1288291	*Anncaliia algerae PRA339*	*Anncaliia*	NA	NA	Microsporidia	Broad
50	278021	*Antonospora locustae*	*Antonospora*	NA	NA	Microsporidia	Broad
51	70536	*Edhazardia aedis*	*Edhazardia*	NA	NA	Microsporidia	Broad
52	6035	*Encephalitozoon cuniculi*	*Encephalitozoon*	NA	NA	Microsporidia	Broad
53	27973	*Encephalitozoon hellem*	*Encephalitozoon*	NA	NA	Microsporidia	Broad
54	58839	*Encephalitozoon intestinalis*	*Encephalitozoon*	NA	NA	Microsporidia	Broad
55	31281	*Enterocytozoon bieneusi*	*Enterocytozoon*	NA	NA	Microsporidia	Broad
56	586133	*Nematocida parisii*	*Nematocida*	NA	NA	Microsporidia	Broad
57	40302	*Nosema ceranae*	*Nosema*	NA	NA	Microsporidia	Broad
58	948595	*Vavraia culicis floridensis*	*Vavraia*	NA	NA	Microsporidia	Broad
59	993615	*Vittaforma corneae ATCC 50505*	*Vittaforma*	NA	NA	Microsporidia	Broad
60	691883	*Fonticula alba*	*Fonticula*	NA	NA	NA	NCBI
61	400682	*Amphimedon queenslandica*	*Amphimedon*	Haplosclerida	Demospongiae	Porifera	JGI

62	45351	*Nematostella vectensis*	*Nematostella*	Actiniaria	Anthozoa	Cnidaria	NCBI
63	946362	*Salpingoeca rosetta*	*Salpingoeca*	Choanoflagellida	NA	NA	NCBI
64	81824	*Monosiga brevicollis*	*Monosiga*	Choanoflagellida	NA	NA	Broad
65	595528	*Capsaspora owczarzaki ATCC 30864*	*Capsaspora*	NA	Ichthyosporea	NA	Broad
66	5762	*Naegleria gruberi*	*Naegleria*	Schizopyrenida	Heterolobosea	NA	JGI
67	5691	*Trypanosoma brucei*	*Trypanosoma*	Kinetoplastida	NA	NA	Sanger
68	5833	*Plasmodium falciparum*	*Plasmodium*	Haemosporida	Aconoidasida	Apicomplexa	plasmodb.org
69	237895	*Cryptosporidium hominis*	*Cryptosporidium*	Eucoccidiorida	Coccidia	Apicomplexa	NCBI
70	3055	*Chlamydomonas reinhardtii*	*Chlamydomonas*	Chlamydomonadales	Chlorophyceae	Chlorophyta	JGI
71	3702	*Arabidopsis thaliana*	*Arabidopsis*	Brassicales	NA	Streptophyta	uniprot
72	67593	*Phytophthora sojae*	*Phytophthora*	Oomycetes	NA	NA	JGI

Table A-3: Mean length of orthologous and orphan proteins in 11 microsporidia.

Taxon	Mean length of orthologous proteins	Mean length of orphans	Wilcoxon-Mann-Whitney P_value
E.hellem	358,507	305,250	0,1966
E.intestinallis	358,931	174,630	9,11E-07
E.cuniculi	368,688	187,100	1,14E-10
N.ceranae	339,184	279,514	2,32E-09
E.bieneusi	274,151	182,634	p < 2,2E-16
V.corneae	330,872	283,743	5,05E-08
A.algerae	284,651	223,355	p < 2,2E-16
A.locustae	295,033	157,594	p < 2,2E-16
E.aedis	380,879	319,525	p < 2,2E-16
V.culicis	370,504	294,433	p < 2,2E-16
N.parisii	421,400	302,794	p < 2,2E-16

Table A-4: GO term annotation for 42 microsporidia specific proteins using Blast2GO.

LCA protein	GO number	Description
OG_1087	P:GO:0017183	P:peptidyl-diphthamide biosynthetic process from peptidyl-histidine
OG_1182	C:GO:0016021	C:integral component of membrane
OG_1323	F:GO:0008080	F:N-acetyltransferase activity
OG_1327	C:GO:0005643	C:nuclear pore
OG_1327	P:GO:0016973	P:poly(A)+ mRNA export from nucleus
OG_1349	C:GO:0016020	C:membrane
OG_1349	P:GO:0016192	P:vesicle-mediated transport
OG_1378	F:GO:0005515	F:protein binding
OG_1515	F:GO:0004672	F:protein kinase activity
OG_1515	F:GO:0005524	F:ATP binding

OG_1515	P:GO:0006468	P:protein phosphorylation
OG_1645	F:GO:0005515	F:protein binding
OG_1649	F:GO:0005524	F:ATP binding
OG_1649	F:GO:0016881	F:acid-amino acid ligase activity
OG_1649	P:GO:0045116	P:protein neddylation
OG_1706	C:GO:0016020	C:membrane
OG_1710	F:GO:0005515	F:protein binding
OG_1731	F:GO:0005515	F:protein binding
OG_1793	F:GO:0005515	F:protein binding
OG_1987	F:GO:0003676	F:nucleic acid binding
OG_2013	F:GO:0015078	F:proton transmembrane transporter activity
OG_2013	P:GO:0015991	P:ATP hydrolysis coupled proton transport
OG_2013	C:GO:0033177	C:proton-transporting two-sector ATPase complex, proton-transporting domain
OG_2013	C:GO:0033179	C:proton-transporting V-type ATPase, V0 domain
OG_2280	F:GO:0005515	F:protein binding
OG_2414	F:GO:0005515	F:protein binding
OG_3062	P:GO:0006364	P:rRNA processing
OG_3062	P:GO:0008033	P:tRNA processing

P: Biological process; C: Cellular component; F: molecular function

Table A-5: Recall, precision and F1-score of HamFAS after excluding annotations from archaea and bacteria reference orthologs in comparison to the original HamFAS, BlastKOALA and KAAS by applying on KO-annotated yeast proteins (set 1).

Approach	HamFAS after filtered	Original HamFAS	BlastKOALA	KAAS
Recall	0.9149	0.9152	0.905	0.931
Precision	0.9867	0.9854	0.979	0.984
F1-score	0.9496	0.9490	0.940	0.957

Table A-6: List of 80 microsporidian core genes with the descriptions from *Saccharomyces cerevisiae*.

No.	ID	Description from *S.cerevisiae*
1	OG_1332	CCR4-NOT core subunit CAF40
2	OG_1336	type 1 serine/threonine-protein phosphatase catalytic subunit GLC7
3	OG_1338	EKC/KEOPS complex
4	OG_1347	ribosomal 60S subunit protein L12A
5	OG_1354	karyopherin alpha
6	OG_1355	ribosome-binding protein NMD3
7	OG_1357	proteasome regulatory particle lid subunit RPN7
8	OG_1358	Chain D, PolyA polymerase module of the cleavage and polyadenylation factor (CPF)
9	OG_1362	rRNA methyltransferase NOP1
10	OG_1364	guanylate kinase
11	OG_1365	replication factor C subunit 5
12	OG_1370	ribosomal 60S subunit protein L13A
13	OG_1374	chromatin-remodeling protein SPT16
14	OG_1375	arginine--tRNA ligase
15	OG_1379	cysteine desulfurase
16	OG_1381	proteasome core particle subunit beta 3
17	OG_1383	cysteine--tRNA ligase
18	OG_1388	ribosomal 40S subunit protein S9B
19	OG_1392	pseudouridine synthase CBF5
20	OG_1394	18S rRNA (guanine1575-N7)-methyltransferase
21	OG_1396	mRNA (guanine-N7)-methyltransferase
22	OG_1400	replication factor A subunit protein RFA1
23	OG_1407	GTPase NPA3
24	OG_1410	AAA family ATPase SEC18
25	OG_1413	tRNA (guanine) methyltransferase
26	OG_1414	ribosomal 60S subunit protein L10

27	OG_1421	ribosome biosynthesis protein KRR1
28	OG_1425	lysine--tRNA ligase KRS1
29	OG_1427	TATA-binding protein
30	OG_1428	ribosomal 60S subunit protein L16A
31	OG_1431	Hsp90 family chaperone HSC82
32	OG_1433	phosphomannomutase SEC53
33	OG_1441	AAA family ATPase midasin
34	OG_1443	diphthine--ammonia ligase
35	OG_1444	proteasome core particle subunit beta 5
36	OG_1445	cleavage polyadenylation factor subunit MPE1
37	OG_1447	proteasome regulatory particle base subunit RPT5
38	OG_1449	type 2A-related serine/threonine-protein phosphatase SIT4
39	OG_1453	ribosomal 60S subunit protein L15A
40	OG_1454	ATP-dependent DNA helicase
41	OG_1459	chaperonin-containing T-complex alpha subunit TCP1
42	OG_1460	ribosomal 40S subunit protein S0A
43	OG_1463	phenylalanine--tRNA ligase subunit beta
44	OG_1470	DEAD-box ATP-dependent RNA helicase DBP2
45	OG_1472	proteasome core particle subunit beta 1
46	OG_1478	AAA family ATPase CDC48
47	OG_1481	GTPase BMS1
48	OG_1486	Rab GDP-dissociation inhibitor
49	OG_1487	glutathione peroxidase GPX2
50	OG_1494	serine/threonine protein kinase KIN2
51	OG_1497	CCR4-NOT core DEDD family RNase subunit POP2
52	OG_1499	actin
53	OG_1501	cyclin-dependent serine/threonine-protein kinase CDC28
54	OG_1503	histone acetyltransferase GCN5
55	OG_1505	exportin CRM1
56	OG_1506	isoleucine--tRNA ligase ILS1
57	OG_1508	syntaxin-binding protein
58	OG_1509	pseudouridine synthase PUS1
59	OG_1511	rRNA (cytosine-C5-)-methyltransferase NOP2
60	OG_1513	Chain A, The Structure Of Glutaminyl-trna Synthetase
61	OG_1514	palmitoyltransferase YKT6
62	OG_1520	replication factor C subunit 3
63	OG_1522	ribosomal 60S subunit protein L7A
64	OG_1524	ribosomal 40S subunit protein S3
65	OG_1530	Ran GTPase GSP1
66	OG_1532	NuA4 histone acetyltransferase complex catalytic subunit ESA1
67	OG_1533	translation termination factor eRF1

68	OG_1536	translation initiation factor 6
69	OG_1538	transferase
70	OG_1539	superoxide dismutase SOD2
71	OG_1540	exosome (RNase complex)
72	OG_1548	RNA-binding signal recognition particle subunit SRP54
73	OG_1550	H(+)-transporting V1 sector ATPase subunit B
74	OG_1551	ribosomal 60S subunit protein L1A
75	OG_1552	DNA topoisomerase 3
76	OG_1553	ribosomal 40S subunit protein S4B
77	OG_1556	TFIIH/NER complex ATP-dependent 5'-3' DNA helicase subunit RAD3
78	OG_1560	DNA-directed RNA polymerase III subunit C34
79	OG_1563	zinc ion binding
80	OG_1573	MCM DNA helicase complex subunit MCM5

Table A-7: Annotated microsporidian proteins for PDH complex, trehalose synthesis and degradation, as well as NTT proteins.

LCA protein	KO id	Description
	Components of pyruvate dehydrogenase complex (PDH)	
OG_2283	K00161	pdhA, E1 component
OG_2084	K00162	pdhB, E1 component
OG_3281	K00382	DLD, E3 component
	Enzymes for trehalose synthesis and degradation	
OG_1266	K00697	trehalose 6-phosphate synthase
OG_1267	K01194	alpha-trehalase
	Nucleotide transport (NTT) proteins	
OG_1062	K03301	NTT1 (Q8SRA2)
OG_1062	K03301	NTT2 (Q8SRA2)
OG_3238	K03301	NTT3 (Q8SUG0)
OG_3237	K03301	NTT4 (Q8SUG7)

Figures

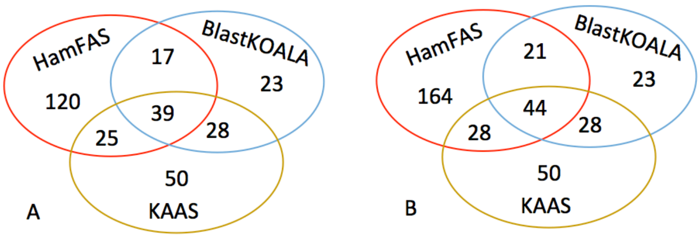

Figure A-1: Fraction of proteins annotated by BlastKOALA, KAAS and HamFAS after excluding annotations from archaea and bacteria reference orthologs (A) or original HamFAS (B).

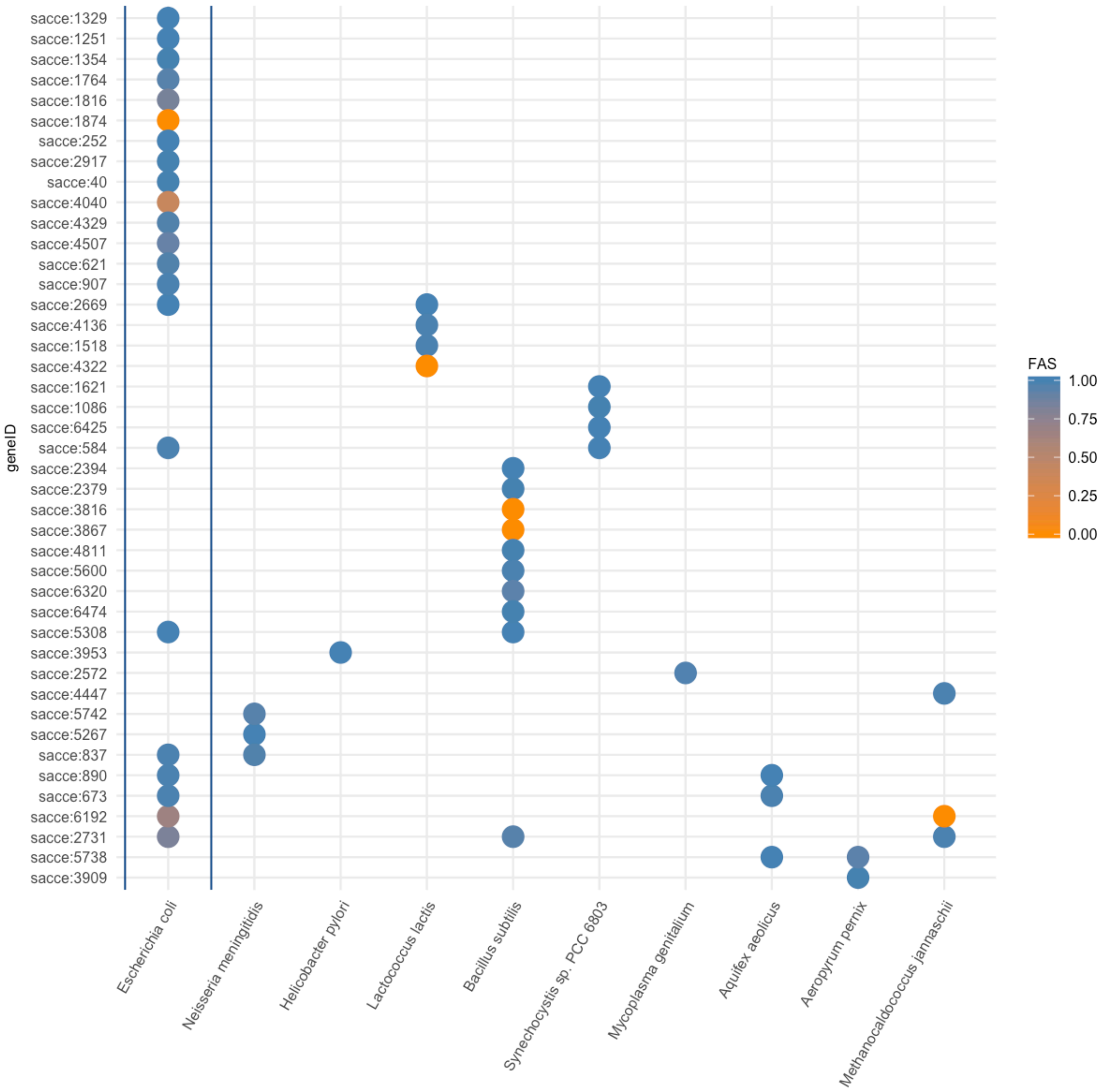

Figure A-2: Phylogenetic profile of 44 HamFAS-only proteins that annotated based on archaea and bacterial orthologs.

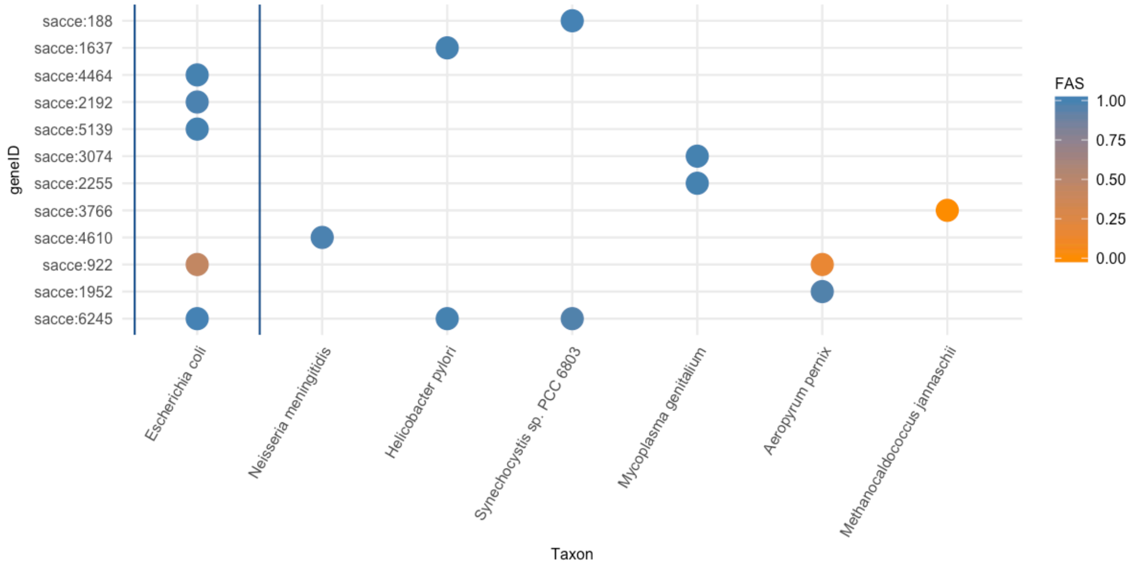

Figure A-3: Phylogenetic profile of 12 un-annotated proteins that annotated by HamFAS and at least one other approach (BlastKOALA and/or KAAS), where their annotations originate from archaea or bacteria reference taxa.

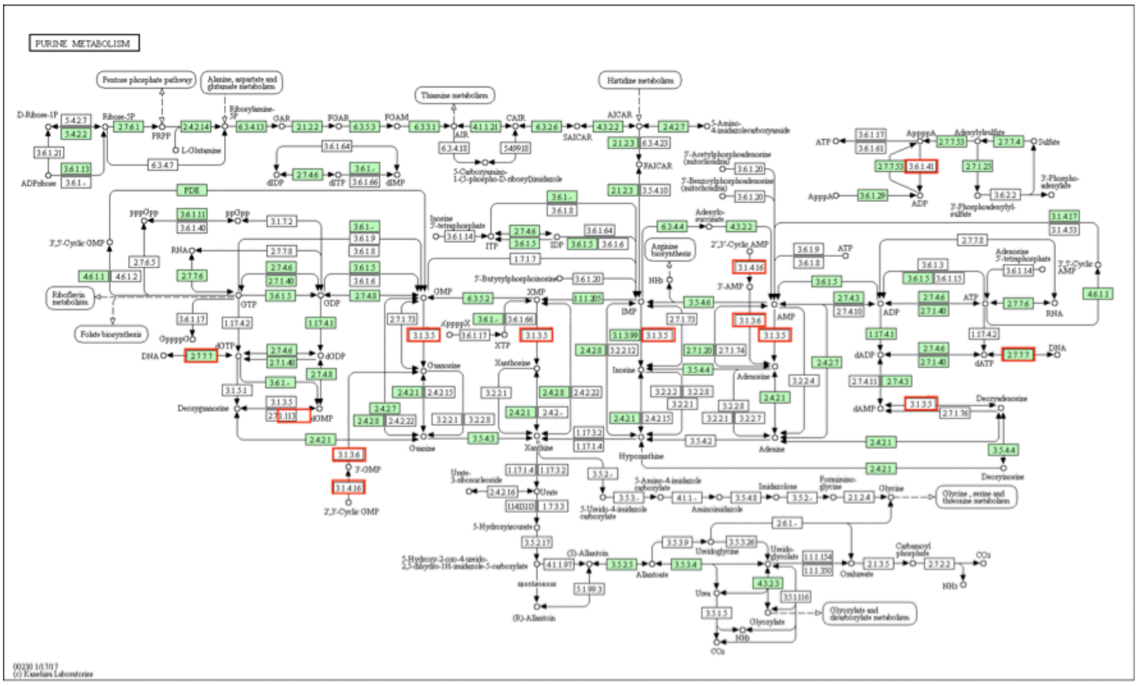

Figure A-4: Purine metabolism for HamFAS annotated yeast proteins. Green highlighted boxes are yeast proteins already present in the KEGG database. Red boxes are complementary proteins from the HamFAS-only annotation. The pathway scheme was obtained from KEGG.

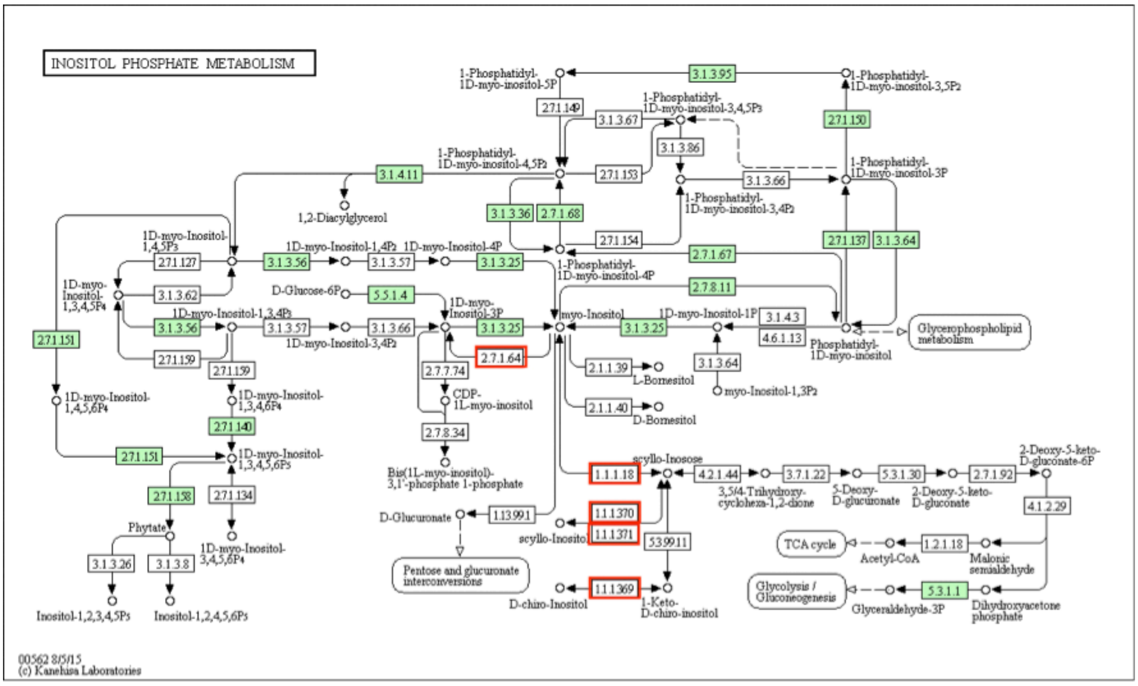

Figure A-5: Inositol phosphate metabolism for HamFAS annotated yeast proteins. Green highlighted boxes are yeast proteins already present in the KEGG database. Red boxes are complementary proteins from the HamFAS-only annotation. The pathway scheme was obtained from KEGG.

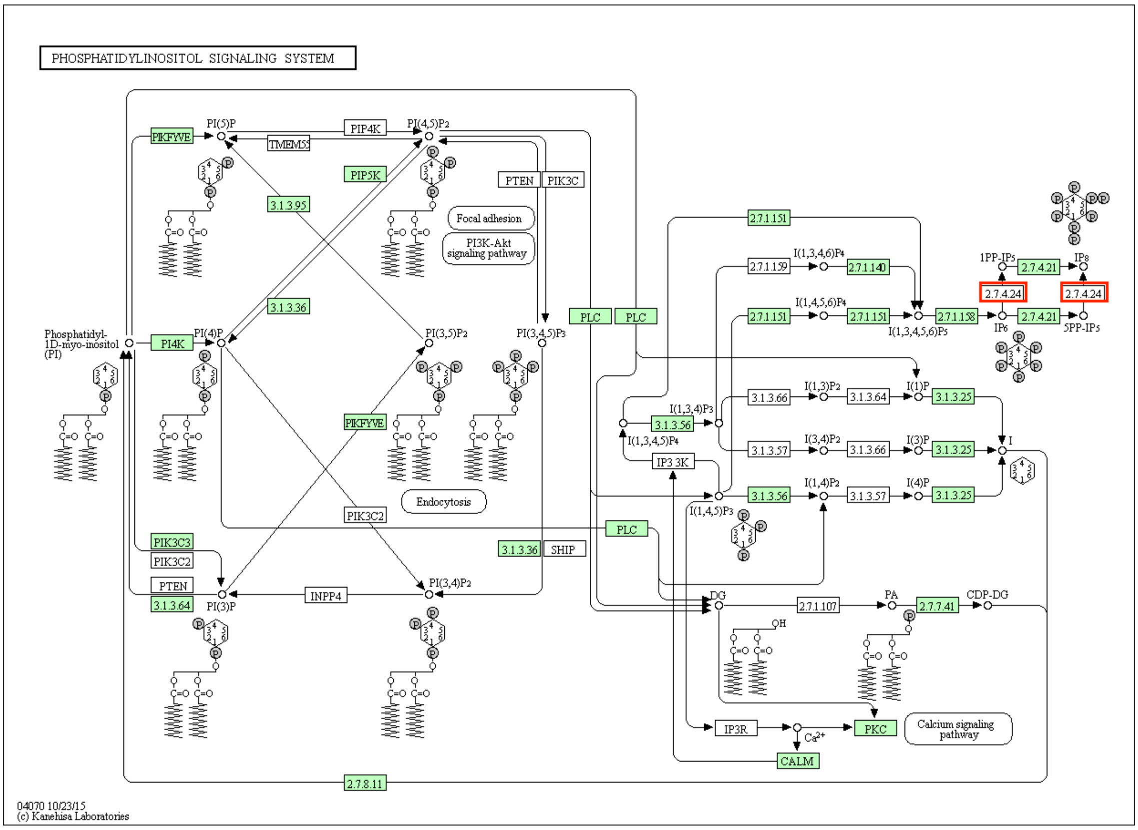

Figure A-6: Phosphatidylinositol signaling system for HamFAS annotated yeast proteins. Green highlighted boxes are yeast proteins already present in the KEGG database. Red boxes are

complementary proteins from the HamFAS-only annotation. The pathway scheme was obtained from KEGG.

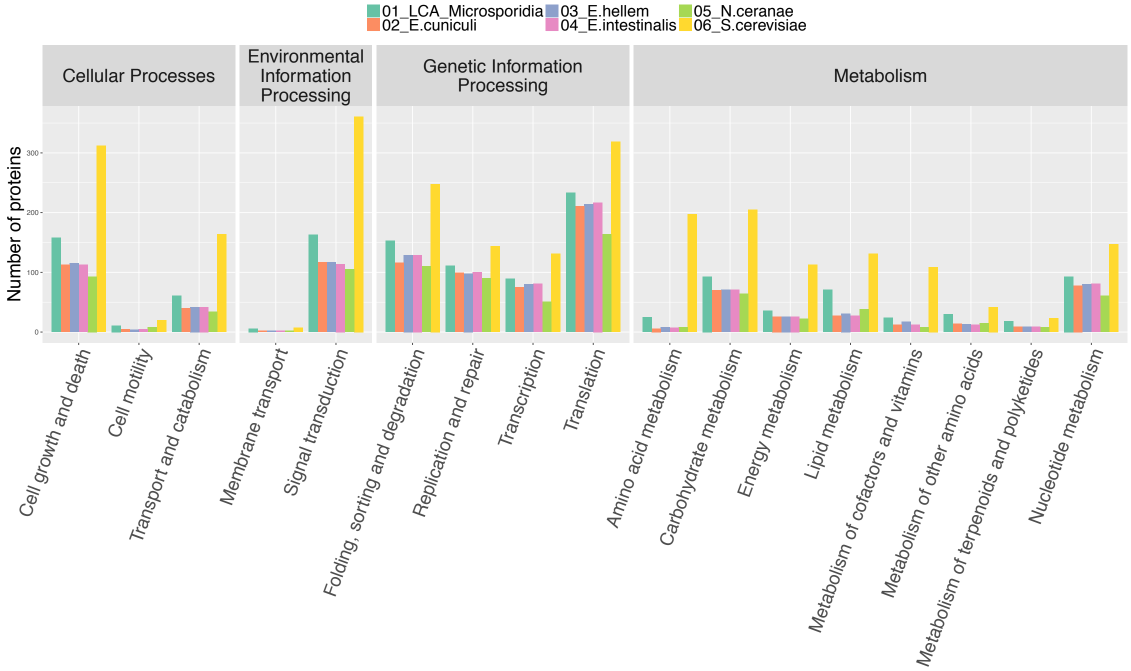

Figure A-7: Number of proteins participating in difference KEGG pathways. Colors denote taxa: dark green for the microsporidian LCA, orange for *E.cuniculi*, purple for *E.hellem*, pink for *E.intestinalis*, light green for *N.ceranae* and yellow for *S.cerevisiae*.

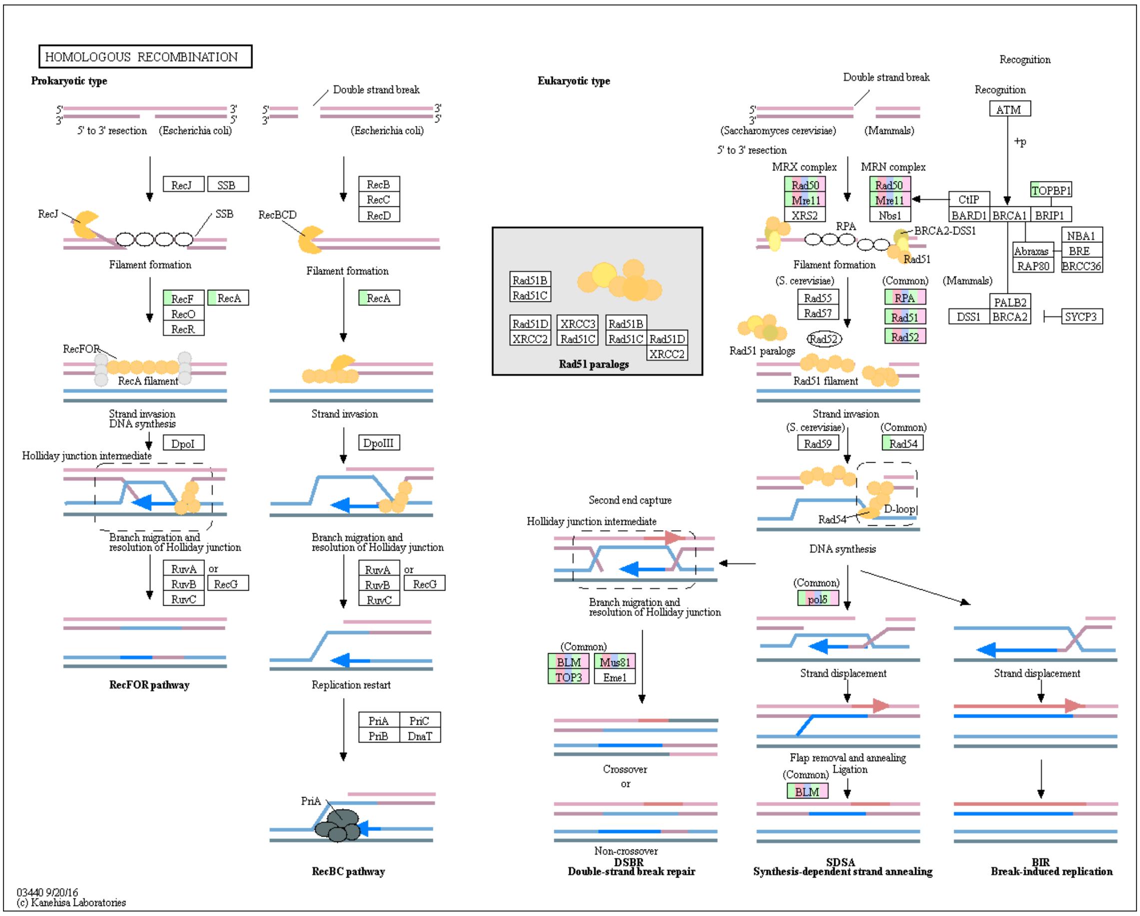

Figure A-8: Scheme of homologous recombination in the microsporidian LCA in comparison to 4 extant species. The mapped proteins are highlighted. The order of the color bars in each annotated proteins is: the microsporidia LCA, *E.cuniculi*, *E.hellem*, *E.intestinalis* and *N.ceranae*. Image obtained from KEGG Mapper.

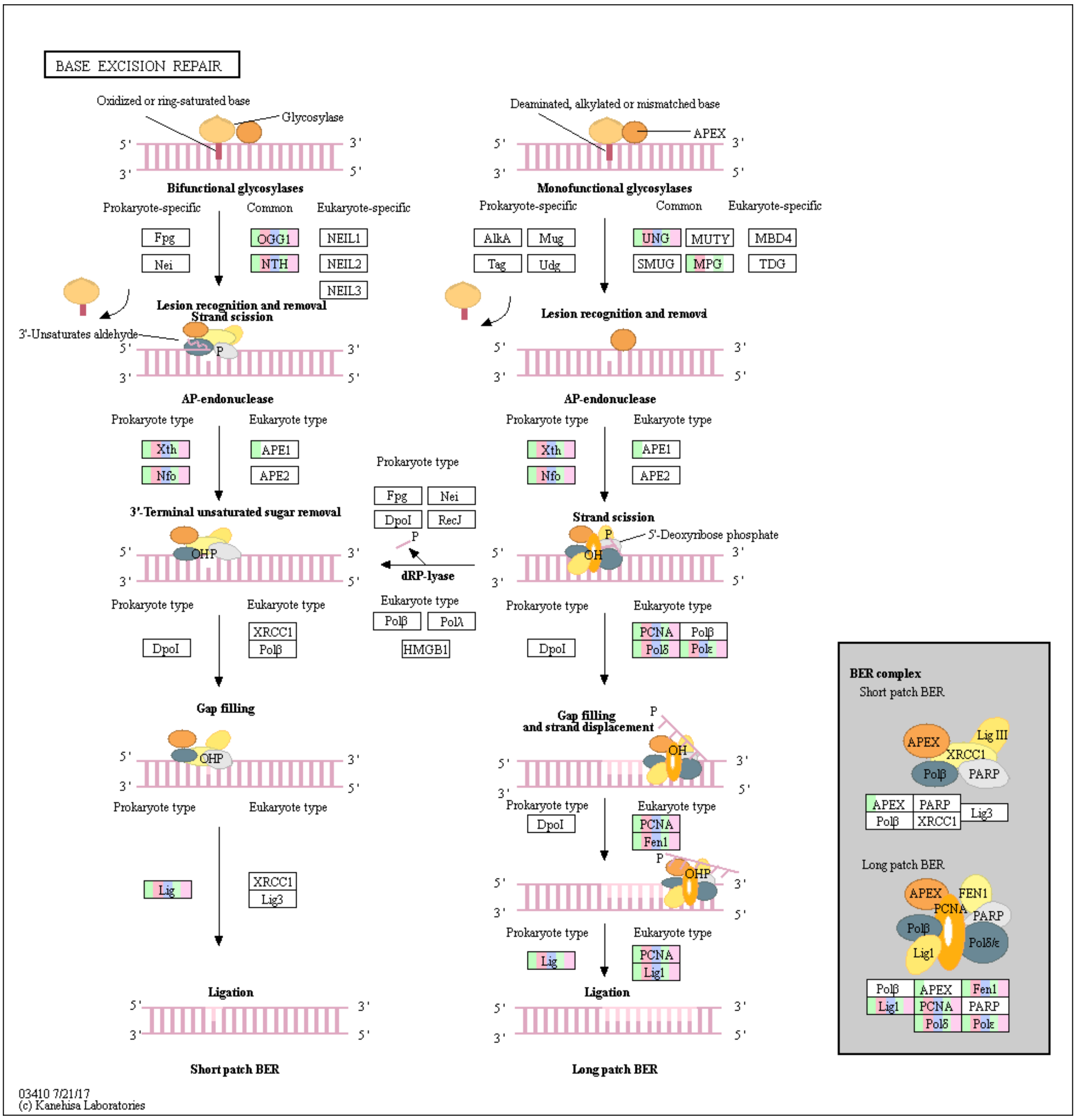

Figure A-9: Scheme of base excision repair process in the microsporidian LCA in comparison to 4 extant species. The mapped proteins are highlighted. The order of the color bars in each annotated proteins is: the microsporidia LCA, *E.cuniculi*, *E.hellem*, *E.intestinalis* and *N.ceranae*. Image obtained from KEGG Mapper.

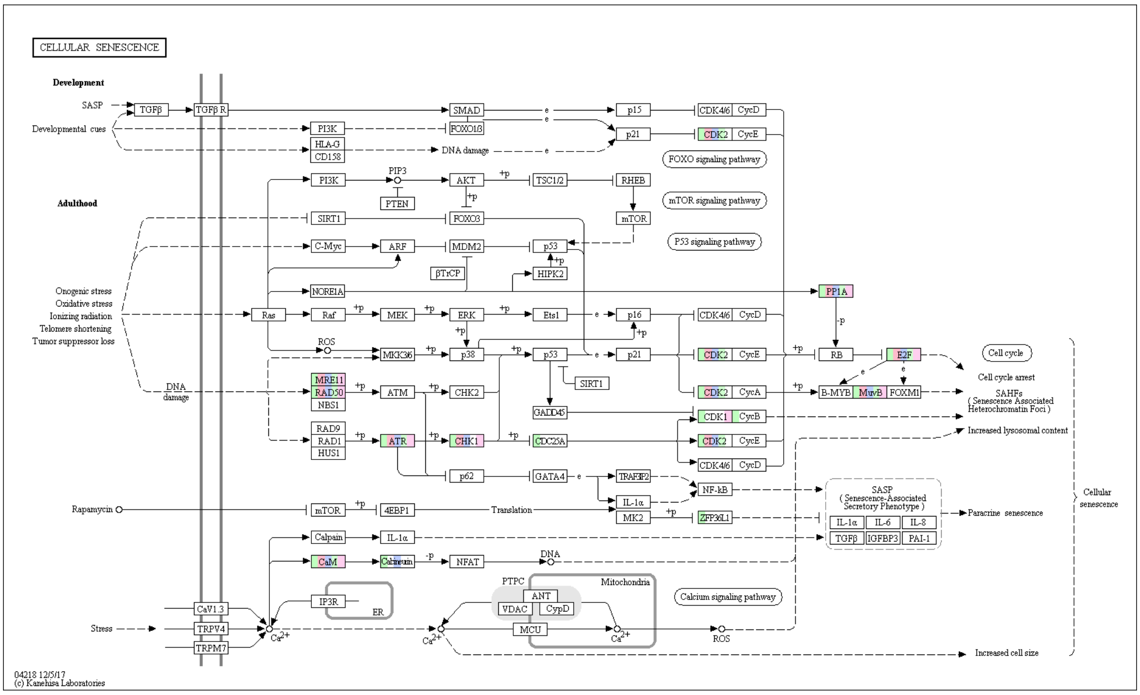

Figure A-10: Scheme of cellular senescence pathway in the microsporidian LCA in comparison to 4 extant species. The mapped proteins are highlighted. The order of the color bars in each annotated proteins is: the microsporidia LCA, *E.cuniculi*, *E.hellem*, *E.intestinalis* and *N.ceranae*. Image obtained from KEGG Mapper.

Acknowledgements

I would never have been able to finish my thesis without the guidance and the advices of several individuals, who in one way or another contributed in the preparation and completion of this study.

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Dr. Ingo Ebersberger for his excellent guidance, patience, and providing me with his broad knowledge. He gave me invaluable help and insightful comments in all the time of my research as well as writing of this thesis. To me, he is not just a mentor but also my best teacher ever.

I also owe a great debt of gratitude to Bastian Greshake Tzovaras, my best colleague, whose outstanding knowledge of everything, inspired and encouraged me to do science and work open. I am very thankful to all his contributions of time, ideas and feedback in my thesis.

Furthermore, I thank to Anne Hänel for her unconditional and amazing assistance with matters inside or outside the office. Her kindly support reduced the stress level of my life. I own Stefan Biermann many thanks for his professional technical support that helped my project run smoothly.

My sincere thanks goes to Arpit Jain, Holger Bergmann, Sachli Zafari, who provided me the beneficial cooperations, valuable discussions and wonderful research ideas. I am grateful to Jan Koch for his assistance he afforded in the correction of the German version of this thesis' summary.

I also thank Laura Kiesewetter, Simonida Zehr, Julian Dosch, Andreas Blaumeiser, Ruben Iruegas, Bardya Djahanshiri and all other members for their true friendship and warm working environment.

I am grateful to Prof. Dr. Enrico Schleiff for agreeing to review my work as the second supervisor and I am appreciatively indebted to him for his very valuable comments on this thesis

Last but not least, I would like to thank my family in Vietnam for their spiritual support during my study. I am especially thankful to my dear wife Bemun and my little son Susu for all their love and encouragement.

Curriculum vitae

PERSONAL INFORMATION

Name	Ngoc Vinh TRAN
Date of birth	20 Aug. 1986
Place of birth	Lam Dong Province, Vietnam
Marital status	Married
Nationality	Vietnamese
Address	Obere Kreuzäckerstraße 37, D-60435 Frankfurt am Main
Telephone	(+49) 151 5628 0251
Email	trvinh@gmail.com - tran@bio.uni-frankfurt.de

EDUCATION

Since 2013 **Goethe University Frankfurt am Main**
PhD in Bioinformatics
- Title: Towards a reconstruction of the microsporidian last common ancestor gene set.
- Supervision: Prof. Dr. Ingo Ebersberger

2010 –2013 **Bielefeld University**
Master of Bioinformatics and Genome Research
- Title: Genome annotation and metabolic pathway reconstruction of *Actinoplanes sp.* SE50/110 based on comparative genome analysis
- Supervision: Dr. Jörn Kalinowski & Dr. Alexander Goesmann

2009 –2010 **Language center, The Philipps-University of Marburg**
DSH course

2004 –2008 **University of Natural Science, Vietnamese National University – HCMC**
Bachelor of Biotechnology, Specialty in Bioinformatics
- Title: Using homology modeling method to predict discontinuous B-cell epitope of matrix protein of H5N1 virus in Vietnam
- Supervision: Dr. Cam Quy Vo & Prof. Dr. Linh Thuoc Tran

EXPERIENCE

RESEARCH SKILL

Bioinformatics
- Data analysis using R, Python and Perl
- Database
- Comparative genomics
- Metabolic network
- Phylogenetics
- Protein structure prediction & protein functional annotation

Molecular biology		• Primer design and gene cloning; PCR, RT-PCR • Expression and purification of proteins; SDS-PAGE, Western-Blot, ELISA, Immunofluorescence

RESEARCH EXPERIENCE

Since 2013 — **Dept. for Applied Bioinformatics, Inst. for Cell Biology and Neuroscience, Goethe University Frankfurt am Main**
PhD Student

2012 – 2013 — **Senior Research Group "Genome Research of Industrial Microorganisms", Center for Biotechnology, Bielefeld University**
Master thesis

2007 – 2009 — **Collaborative Bioinformatics Laboratory, Department of Molecular Biotechnology and Environment, University of Natural Sciences, HCMC**
Academic staff

TEACHING EXPERIENCE

2013 – Present — **Dept. for Applied Bioinformatics, Inst. for Cell Biology and Neuroscience, Goethe University Frankfurt am Main**
Teaching assistant
Principle of Bioinformatics, Molecular Evolution and Bioinformatics (Practical course)

2008 – 2009 — **Department of Molecular Biotechnology and Environment, University of Natural Sciences, HCMC**
Teaching assistant
Practical course in Bioinformatics

PUBLICATIONS & CONFERENCES

PUBLICATIONS

Ngoc-Vinh Tran, Bastian Greshake Tzovaras, and Ingo Ebersberger (2018), *"PhyloProfile: Dynamic visualization and exploration of multi-layered phylogenetic profiles"*, Bioinformatics, doi: 10.1093/bioinformatics/bty225.

Ana I. S. Moretti, Jessyca C. Pavanelli, Patrícia Nolasco, Matthias S. Leisegang, Leonardo Y. Tanaka, Carolina G. Fernandes, João Wosniak Jr, Daniela Kajihara, Matheus H. Dias, Denise C. Fernandes, Hanjoong Jo, **Ngoc-Vinh Tran**, Ingo Ebersberger, Ralf P. Brandes, Diego Bonatto and Francisco R. M. Laurindo (2017), *"Conserved gene microsynteny unveils functional interaction between protein disulfide isomerase and Rho guanine-dissociation inhibitor families"*, Scientific Reports, vol.7, doi: 10.1038/s41598-017-16947-5.

Tran Ngoc Vinh, Vo Cam Quy and Tran Linh Thuoc (2009), *"Discontinuous B-cell epitope prediction of matrix protein of H5N1 virus"*, Science & Technology Development Magazine, HCMC, Vietnam, vol.12, pp 31-38.

CONFERENCES

Ngoc-Vinh Tran, Bastian Greshake Tzovaras, and Ingo Ebersberger (2017), *"PhyloProfile: Dynamic visualization and exploration of multi-layered phylogenetic profiles."*, Poster at the 18th Annual Bioinformatics Open Source Conference BOSC2017, doi: 10.7490/f1000research.1114937.1.

Holger Bergmann, **Ngoc-Vinh Tran**, Bastian Greshake Tzovaras, Julian Dosch, Bardya Djahanschiri, Sachli Zafari and Ingo Ebersberger (2017), *"Tracing functional protein interaction networks using a feature-aware phyletic profiling"*, Poster at the 2017 SMBE Meeting.

Tran Ngoc Vinh, Vo Cam Quy and Tran Linh Thuoc (2008), *"Discontinuous B-cell epitope prediction of matrix protein of H5N1 virus for vaccine development in silico"*, Proceeding of The 6th Scientific Conference at the University of Natural Sciences, HCMC, Vietnam.

Vo Cam Quy, Nguyen Thi Truc Minh, Nguyen Duc Duy, **Tran Ngoc Vinh** and Tran Linh Thuoc (2008), *"Establishing Influenza A VIrus Database to assist the epitope prediction"*, Proceeding of The 4th National Scientific Conference "Biochemistry and Molecular Biology in the service of agriculture, medicine, biology and food industry" in Hanoi, Vietnam, pp 522-525.