Introduction

The microsporidia and their impact on the economy and human health

Microsporidia are well-known as an unusual group of obligate intracellular parasites. Currently there are approximate 1,400 species have been reported (Dean *et al.*, 2016), which infect a large variety of hosts from vertebrates to invertebrates (Vossbrinck *et al.*, 1987; Scanlon *et al.*, 2000). Microsporidia were soon discovered as pathogenic factors that are responsible for many diseases. The first microsporidia *Nosema bombycis* has been found to be the causative agent for the silkworm disease (pébrine), which has seriously affected the silk industry in the mid-nineteenth century (Vivarès and Méténier, 2001). Other related species, *Nosema apis* and *Nosema ceranae*, cause nosemosis disease on the European honeybee *Apis mellifera* that influenced the commercial honey producers in recent years (Charbonneau *et al.*, 2016). The finfish aquaculture has also been damaged by *Pseudoloma neurophilia* or a number of microsporidia in the genus *Glugea* (Ramsay *et al.*, 2009; Ryan and Kohler, 2016). The first mammalian infection was caused by *Nosema cuniculi* in 1922 (which was reassigned to *Encephalitozoon* in 1923) in brain, spinal cords and kidneys of rabbits (Vivarès and Méténier, 2001). After being detected in human the first time in 1959, microsporidian infections got more attentions when *Enterocytozoon bieneusi* and other species in the genus *Encephalitozoon* were found in the immunocompromised patients (Scanlon *et al.*, 2000; Vivarès and Méténier, 2001). Until now, there are 13 microsporidia species have been reported to be involved in different human diseases (Keeling and Fast, 2002).

The symbiotic lifestyle of microsporidia

A symbiotic relationship is the association between two different organisms, symbionts, in which one symbiont can live inside (endosymbiosis) or outside (ectosymbiosis) the other (Paracer and Ahmadjian, 2000). In the three types of symbiosis, parasitism is a relationship where one symbiont, the parasite, benefits from its partner, the host, by using the resource from the host (Paracer and Ahmadjian, 2000).

As an extreme case of the parasitic relationship, the unicellular microsporidia are obligated dependent on their host (Agnew *et al.*, 2003). Outside the host-cellular environment, microsporidia can only be visible as a spore in different forms with the size range from 1μm to 40μm (Keeling and Fast, 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 *et al.*, 2016).

The reduction of microsporidian genomes and metabolism

The microsporidian genome size is spread from 2.3 Mbp to 23 Mbp (Keeling and Fast, 2002; Peyretaillade *et al.*, 2012). Although they have eukaryotic characteristics such as multiple linear chromosomes or telomeres, their small size makes microsporidia become a model organism for studying reduction in eukaryotic genomes and metabolomes (Keeling and Fast, 2002; Wiredu Boakye *et al.*, 2017). In some cases, microsporidia genome size is just in the range of bacterial intracellular parasites (Vivarès and Méténier, 2001). The species *Enzephalitozoon intestinalis*, whose 2.3 Mbp genome is just half the size of the Escherichia coli's one, is known as the smallest eukaryote species (Corradi *et al.*, 2010). As a result of the genome reduction, the microsporidia have only approximately 1,750 to 3,266 protein coding genes, which are thought to be essential for their parasitic survival (Agnew *et al.*, 2003; Nakjang *et al.*, 2013). The microsporidian genes are mostly shorter than their orthologs from other organisms (Katinka *et al.*, 2001). They are flanked by short intergenic spaces, have few introns and repeat sequences (Keeling and Fast, 2002; Corradi *et al.*, 2010). Moreover, some of the genes are overlapped with each others (Corradi *et al.*, 2010).

The compact genome of microsporidia is the evolutionary result of the prosperous adaption to their obligate intracellular parasitic lifestyle (Nakjang *et al.*, 2013). Microsporidia are strongly depend on their host for nutrients due to the lack of mitochondria and genes for many biosynthesis pathways (Katinka *et al.*, 2001; Luallen *et al.*, 2016). They produce ATP through glycolysis instead of the more efficient Kreb cycle, or use this energy resource from their host species via the transport system (Dolgikh, 2000; Keeling and Corradi, 2011; Heinz *et al.*, 2012). Some nucleotide transport proteins have been found in microsporidia to replace the missing of ribose-phosphate pyrophosphokinase and IMP cyclohydrolase genes, which are essential for the *de-novo* purine and pyrimidine biosynthesis (Heinz *et al.*, 2014; Dean *et al.*, 2016).

The origin of microsporidia

Initially, the microsporidia *Nosema bombycis* was described as a yeast-like unicellular fungus by (Naegeli, 1857) . Thereafter, some electron microscopy studies reassigned microsporidia in the phylum Sporozoa, and then together with other amitochondriate protists in the phylum Archezoa because they lacked several typical eukaryotic components such as mitochondria, Golgi bodies or peroxisomes (Kudo R. R. and Daniels E. W., 1963; Heinz *et al.*, 2014). The first molecular based phylogenetic analysis from the small and large subunit of ribosomal RNAs of the microsporida *Vairimorpha necatrix* further approved the Archezoa hypothesis (Vossbrinck *et al.*, 1987). Since then, this earliest eukaryotic origin of microsporidia further supported with the phylogeny of other genes such as isoleucyl aminoacyl-tRNA synthetase, elongation factor-1alpha, and elongation factor-2 (Corradi and Keeling, 2009). However, this ‘‘Microsporidia-early’’ hypothesis was always doubted. The fast evolving or divergent sequences of the microsporidia could mislead their deep position in the phylogenetic tree due to the effect of the long-branch attraction (Felsenstein, 1978; Corradi and Keeling, 2009). After more than 100 years from the report of Naegeli, microsporidia were re-classified as fungi by placing them either within or in the earliest branch of the fungal clade (Cavalier-Smith, 2004). This fungal relationship was firstly proved by the phylogenetic study of alpha- and beta-tubulins from several microsporidia species (Keeling and Doolittle, 1996; Edlind *et al.*, 1996). Afterward, the evidence for this hypothesis has been becoming stronger with the analyses of the heat-shock protein 70 (Hirt *et al.*, 1997), the largest subunit of the RNA polymerase II (Hirt *et al.*, 1999) ,both alpha and beta subunits of pyruvate dehydrogenase E1 (Fast and Keeling, 2001), the new DNA-dependent RNA polymerase II largest subunit RPB1 and translation elongation factor I alpha (Tanabe *et al.*, 2002), or the combined approach using 53 different genes of (Capella-Gutiérrez *et al.*, 2012). However, those researches were not agreed on an exact relationship between microsporidia and fungi, if microsporidia are related to the class Ascomycetes, Basidiomycetes, Zygomycetes, phylum Cryptomycota or they are the sister group of fungi (Keeling and Fast, 2002; Koestler and Ebersberger, 2011; Heinz *et al.*, 2012; James *et al.*, 2013), mostly due to the poor data sampling (Capella-Gutiérrez *et al.*, 2012).

Potential research of microsporidia

With the economic and clinical importance, microsporidia become an interesting subject for studying the parasitic lifestyle in eukaryote and the microsporidian pathobiology. However, the *in vitro* experiments on microsporidia still have limitations. Since these species obligatory depend on the host cells for their development, the purified samples can contain only the microsporidian spores (Méténier and Vivarès, 2001). Nevertheless, the physiology of the sporal stage are thought to be different from the developmental stages inside the host cell (Dolgikh *et al.*, 1997). Therefore, just a limited knowledge about the microsporidian genomic evolution and biochemical metabolic pathways were achieved (Heinz *et al.*, 2012; Nakjang *et al.*, 2013). For a better understanding about the biological interaction between the microsporidia and the host species as well as their position in the tree of life, it is necessary to investigate the ancestor of these eukaryotic parasites (Keeling and Fast, 2002).

Thus, we carried out this study to examine the fungal related origin of microsporidia and to explore the metabolic network of their last common ancestor. In Chapter 2, " The estimation of the microsporidian last common ancestor protein set ", we used a parsimony approach to estimate the protein set of microsporidia last common ancestor (LCA), which was then the basic data for the downstream analyses. In order to analyze the phylogenetic distribution of the microsporidian LCA proteins in an effective and informative manner, we created an phylogenetic visualization and analysis tool named PhyloProfile, which is introduced in Chapter 3, "PhyloProfile: an interactive visualization tool for exploring complex phylogenetic profiles". In Chapter 4, "Distribution analysis of microsporidian LCA proteins", we applied PhyloProfile to the protein set of the microsporidian LCA to measure the evolutionary ages of those sequences. Recognized the influence of the orthology assignment to the protein functional annotation, we developed a novel functional annotation approach HamFAS, which utilize both evolutionary relationship and domain structure information to transfer protein annotations. This approach is described in Chapter 5, " HamFAS: a novel functional annotation approach based on feature-aware orthology inference ". In the last chapter, " Metabolic pathway analysis of the microsporidian LCA proteins", we applied HamFAS to assign the functional annotations to the microsporidia LCA proteins and based on those annotations, the microsporidian LCA metabolic network was compared with those of the extant species. (ADD HYPERLINK TO CHAPTERS)

The estimation of the microsporidian last common ancestor protein set

Introduction

The analysis of species phylogeny can give insight into the evolutionary history of those species, such as what is the systematic relationship between them and the others in the phylogenetic tree of life, or how their pathways evolved across species 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). For this reason, in this chapter we describe an orthology-based approach for estimating the microsporidian last common ancestor (LCA) protein set, which was served as an initial data for further analyses.

Methods

Data collection

In the scope of this study, we used an representative data set comprise of eleven microsporidia species downloaded from the genome portal of the Join Genome Institute (Nordberg *et al.*, 2014) and 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). Those selected taxa includes species with different size of proteomes, from the most compact *Encephalitozoon intestinalis* with 1657 proteins to the larger *Edhazardia aedis* with 4208 proteins. Details about the data set can be found in Table A‑1.

Orthology prediction

First, we used OrthoMCL (Li *et al.*, 2003) to search for homologous proteins within the representative microsporidia taxon set. OrthoMCL performed all-against-all BLASTP comparisons for all input data set and clustered homologous groups together using the Markov Cluster algorithm MCL (van Dongen, 2000).

Then, we extended the homologous groups retrieved from OrthoMCL by searching for their orthologs in other 24 search taxa (Table A‑2) using HaMStR approach (Ebersberger *et al.*, 2009). HaMStR uses the Hidden Markov Model (HMM) profiles (Eddy, 1998) from the initial homologous groups (seed sequences) to search for the similar sequences in the search taxa. The obtained best 10 HMM hits were confirmed by a reverse BLAST search (Altschul *et al.*, 1990) against the protein sets of seed species. As microsporidia genes tend to evolve quickly (Lee *et al.*, 2008), the BLAST search could be false to return the seed sequence as its best hit. We therefore increased the sensitivity of the prediction by accepting the seed protein to be co-orthologous to the best reverse BLAST hit.

Species tree reconstruction

After having the extended orthologous groups, we identified a core gene set to reconstruct the maximum likelihood species tree of the 35 selected taxa. We defined core genes as orthologous groups that contain orthologs in all taxa and each taxon is presented by exactly one protein. Firstly, we aligned the sequences of individual orthologous groups of the core genes with the program ClustalW (Larkin *et al.*, 2007). Secondly, a super-alignment was generated by concatenating those single alignments together. As proposed from other microsporidian phylogenetic study to reduce the influence of the long branch attraction (Keeling and Fast, 2002; James *et al.*, 2013), we kept only alignment columns that have less then 50% of gaps. Then, we used ProtTest (Abascal *et al.*, 2005) to find the best fitting model for the tree reconstruction procedure using the de-gapped super-alignment. At last, we reconstructed the maximum likelihood species tree from the processed super-alignment using the tool RAxML (Stamatakis, 2014). The tree reconstruction was done based on the best model parameters obtained from ProtTest with 100 bootstrap replicates.

Last common ancestor's proteins estimation

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 representing in the microsporidian LCA. Those final orthologous groups must have either (1) at least one ortholog from *Nematocida parisii* (the earliest branch in the microsporidia lineage shown in the reconstructed tree), or (2) at least two orthologs from other microsporidia species different than *Nematocida parisii* and one or more orthologs from non-microsporidia taxa.

Results

Using OrthoMCL, we obtained 2904 initial homologous groups for eleven microsporidia taxa. Because of the methodology used in OrthoMCL, those groups contain both orthologs between species and in-paralogs within the species. We kept all of them for the extension ortholog search with HaMStR.

Out of 2904 extended groups, we identified 80 orthologous groups, where all 11 microsporidia and 24 non-microsporidia taxa are present and each taxon has one representative ortholog. Those 80 groups served as our core genes for the species tree reconstruction.

The super-alignment concatenated from 80 single alignments of the core gene set has the length of 86.424 positions. After removing columns that have at least 50% gaps, we got a final super-alignment of length 36.616.

The best fitting model for that de-gapped super-alignment retrieved from ProtTest was LG substitution model (Le and Gascuel, 2008), GAMMA distribution G , including proportion of invariable sites estimation I (Steel *et al.*, 2000) & empirical base frequencies F.

We performed the species tree reconstruction with RAxML using the de-gapped super-alignment and input model parameter PROTGAMMAILGF in 100 bootstrap replicates. The reconstructed maximum likelihood tree with bootstrap support values is shown in Figure A‑1 below.

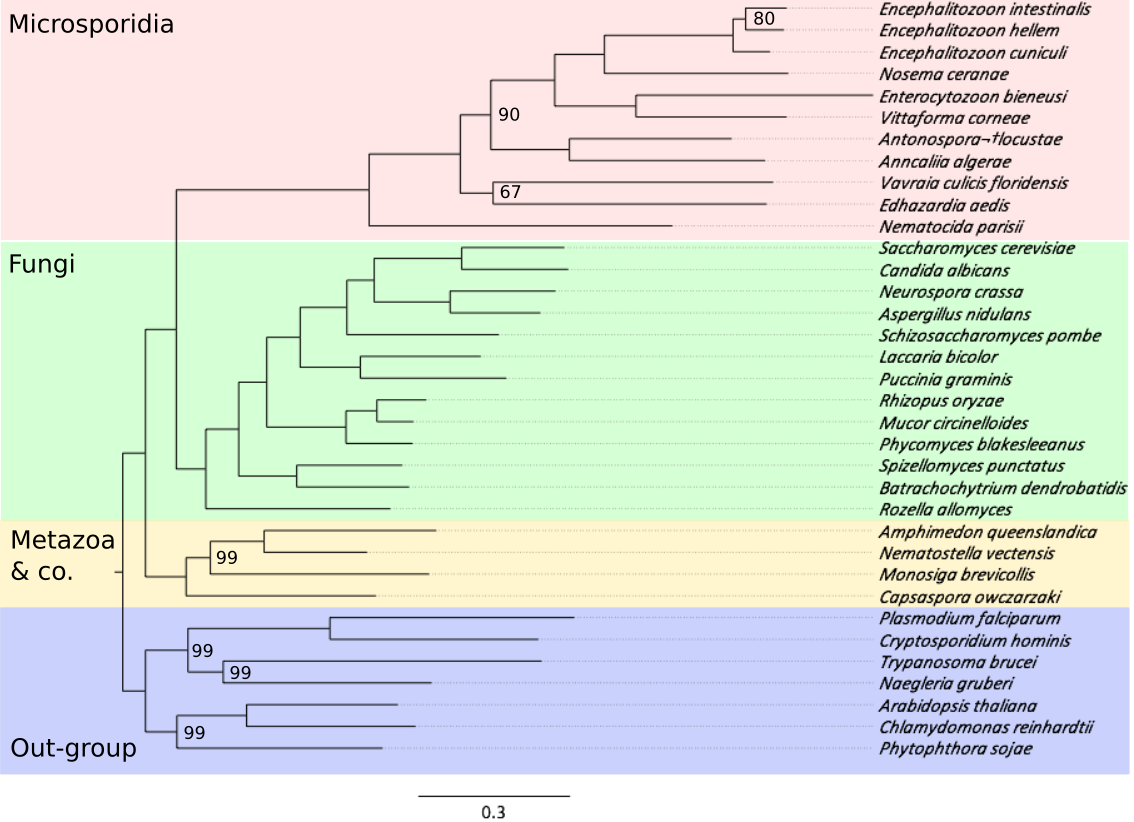


Figure A‑1: Maximum likelihood tree over 35 species. The 11 microsporidia taxa are highlighted in red. Other non-microsporidia taxa include 13 Fungi (green), 2 Metazoa and M.brevicollis, C.owczarzaki (yellow) and 7 out-group species (purple). InternalNode labels denote the bootstrap support and only values less than 100 are shown.

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

Discussion

Proportion of orthologous and lineage specific proteins

We analyzed the proportion of orthologous and lineage specific proteins in 11 microsporidia species (Figure A‑2).

As a model for the genome reduction in microsporidia, the species in *Encephalitozoon* genus share almost 98% of their proteins with other microsporidia species. While only 2% are lineage specific proteins (orphans). In other taxa, there are still a large fraction of orphan proteins, from approximately 21% in *N.ceranae* up to 49% in *E.aedis*.

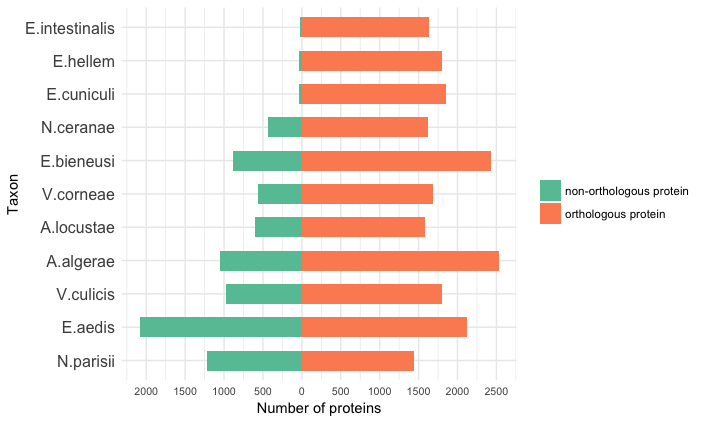


Figure A‑2: Fractions of non-orthologous (orange) and orthologous (green) proteins in different microsporidia species.

We investigated the orphan proteins with the following hypotheses.

(1) Wrong gene assignment:

In this case, we hypothesize that those orphans were the false positive result of the gene prediction process. To access this, we compared the sequence length of orphans and orthologous proteins with the assumption that wrongly assigned genes would be shorter than the corrected one.

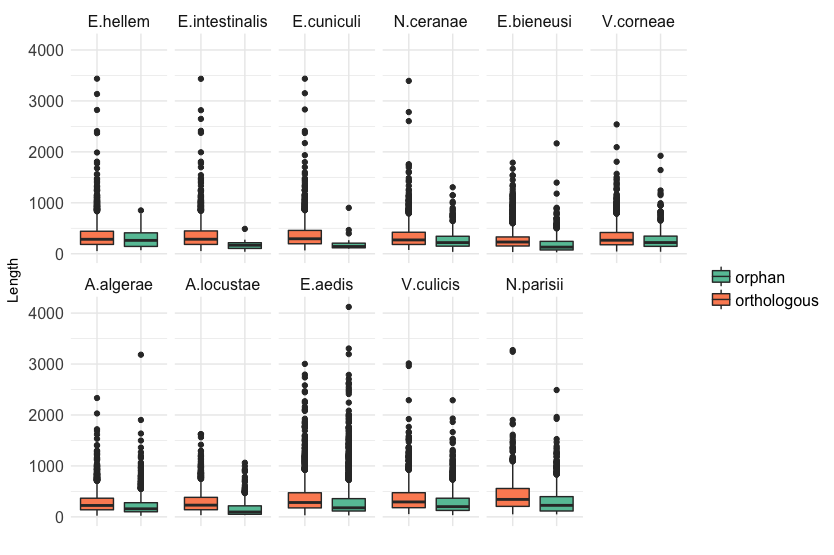


Figure A‑3: Length distribution of orthologous proteins (orange) and orphan proteins (green) in different microsporidia taxa.

Figure A‑3 shows the length distribution of orthologous and orphan proteins in 11 microsporidia taxa. We use the nonparametric U-test Wilcoxon-Mann-Whitney (Mann and Whitney, 1947) to compare the two length categories. We found that the lengths of orphan proteins are significantly smaller than the one of orthologous proteins with the significant level of 5%. Only in the case of *E.hellem*, the p-value was 0,20 > 0,05. However, the comparison in cases of *Encephalitozoon* taxa should be ignored because of the small number of orphan proteins that leads to a huge difference in sample size between the two datasets, which could shrink the power of the U-test (Noether, 1987).

(2) Orphans are new invented genes, or genes from horizontal gene transfer events, or they cannot be detectable as orthologs.

If the orphan genes are correctly predicted, there are some possible explanations for the missing of their orthologous partners in other species. To assess those assumptions, we performed a PFAM (Finn *et al.*, 2014) annotation analysis for the orphan and orthologous proteins in each microsporidia species. The PFAM annotation was done using hmmscan (Eddy, 1998) to search for the similar sequences from the pfam-A database.

Firstly, the orphans could be new genes, which have been invented after the speciation event within the microsporidia lineages. For those new acquired genes we expected to find no PFAM domain. Secondly, the orphans could be horizontal transferred from other taxa. In this case, they would have non-microsporidian PFAM domains, which cannot be found in the orthologous proteins. Lastly, those orphan genes have been evolved quickly so that we could not find their orthologous partners with the current sequence similarity based approaches. However, as PFAM domains are the highly conserved regions throughout sequences and species, we expected to find the domains of microsporidian orthologous genes even in the fast-evolved orphans.

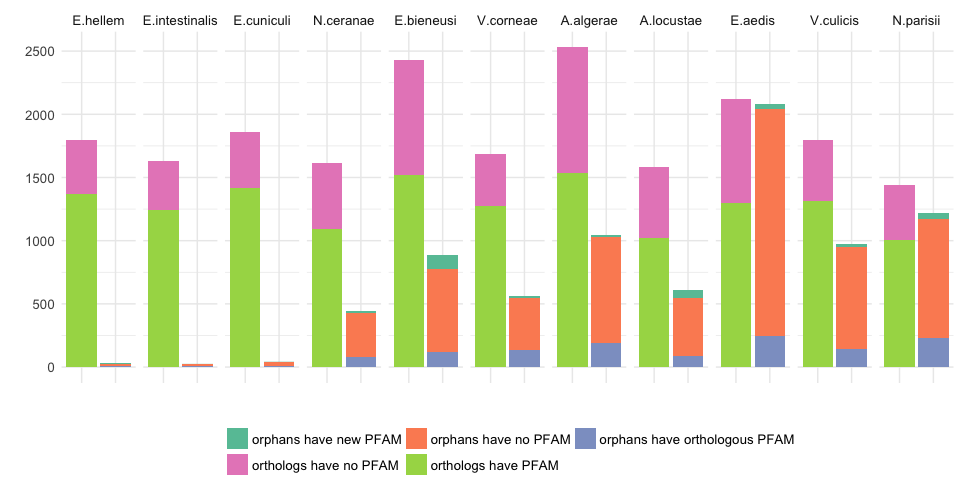


Figure A‑4: Fractions of orthologous and orphan proteins that have and do not have PFAM annotations. The left columns show the number of orthologous proteins that have PFAM annotations (light green) and do not have PFAM annotations (pink). The right columns show the proportion of orphan proteins that have new PFAM annotations that are not found in orthologous proteins (dark green), do not have any PFAM annotation (orange) and orphans that have the same PFAM annotations as orthologous proteins (purple).

A large fraction of orphan proteins (from 70% to 86%) do not have any PFAM domain as been shown in Figure A‑4 suggests that most of those orphan proteins are newly invented during the expansion process of microsporidia genomes. About 12% to 28% of orphan proteins have the same PFAM domains with those from orthologous proteins. According to our assumptions, they could be the fast-evolved proteins that we could not find their orthologs using OrthoMCL and HaMStR approaches.

The core gene set and the microsporidian origin

The identified 80 core genes, which are used for reconstructing the species tree in Figure A‑1, was proved to be a very exceptional set for studying the evolutionary of fungal or even eukaryotic lineage. It was used to investigate the co-evolution of PDI/RhoGID gene clusters (protein disulfide isomerases and Rho guanine-dissociation inhibitors) across the animal phylogeny in the research of (Moretti *et al.*, 2017), or to study the fungal diversity using a large taxon set with 48 fungi including Ascomycoty, Basidiomycota, Blastocladiomycota, Chytridiomycota, Entomophthoromycota, Glomeromycota, Neocallimastigomycota, Kickxellales, Mortierellales and Mucorales, together with 11 microsporidia and 13 other taxa (Figure A‑5).



Figure A‑5: (REMOVE BOOTSTRAP VALUES OF 100) The maximum likelihood fungal tree generated based on the microsporidian core gene set. The tree reconstruction pipeline is similar to the one that was explained in the methods part (point ??.??.??)

Both reconstructed species trees in Figure A‑1 and Figure A‑5 solidly support the hypothesis that microsporidia forms the earliest diverging clade of fungi. It is also worthwhile to mention that the topology of the tree in Figure A‑1 is congruent with the one from the study of (Capella-Gutiérrez *et al.*, 2012), which also support the same scenario.

Conclusion

The estimation of microsporidian LCA proteins is the basic step for the whole downstream analysis. By including eleven microsporidia species, whose public sequences were available at the time of collecting, we expected to have a sufficient taxon sampling for this comparative study.

It has been shown that, even with the intense genome reduction, microsporidia species still have a fraction of lineage specific genes (Peyretaillade *et al.*, 2012). We found 21% to 49% of the assigned proteins were the microsporidian orphan proteins. Excluding the false predicted proteins, most of the orphans are new newly invented genes in the microsporidia clade and some of them are the quickly evolved genes that cannot traceable. 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).

The phylogenetic trees reconstructed from the microsporidia core gene set strongly support the hypothesis that microsporidia is the sister clade of fungi. With a larger taxon sample including more diverge taxa, our data can clarify this assumption better than the one used by (Capella-Gutiérrez *et al.*, 2012).

The microsporidia species tree can serve as a fundamental phylogenetic background for filtering the orthology assignment and estimating the set of 1605 proteins for the LCA of the microsporidia.

PhyloProfile: an interactive visualization tool for exploring complex phylogenetic profiles

Introduction

In evolutionary biology, the presence/absence pattern of a gene 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 *et al.*, 2007; Date and Peregrín-Alvarez, 2008). Moreover, phylogenetic profiles are commonly used for tracing gene clusters or biological pathways across species and time (Li *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 similarity, domain architecture similarity or semantic similarity of Gene Ontology-terms (Kensche *et al.*, 2008).

Currently, there are resources and tools available for such enriched phylogenetic profiles, such as DoMosaics (Moore *et al.*, 2014), the ETE3 tool kit (Huerta-Cepas *et al.*, 2016, 3) or the recently published Aquerium (Adebali and Zhulin, 2017) that are able to display protein domain architectures along a phylogenetic tree. Though, those tools lack a set of comprehensive analysis functions as well as the ability to intensively visualization of multi-layered phylogenetic profiles containing hundreds or thousands of genes and taxa. Hence, we developed PhyloProfile, an interactive visualization tool for dynamically exploring such complex phylogenetic profiles.

Features and capabilities

Interactive visualization

PhyloProfile was written mainly in R (R Development Core Team, 2011). Because of the robust ability of interactive visualization in analyzing informative data (Zudilova-Seinstra *et al.*, 2009), we intensively used the Shiny library (https://CRAN.R-project.org/package=shiny) to brought this feature into PhyloProfile tool.

dynamic data analysis

filtering

link between different data types

change systematic rank to change the scale of the analysis

phylogenetic profiling:

analysis functions

Interoperable output

Result

The availability

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). 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 ......, which is suitable for small to medium-scales analyses or for testing purpose.

Performance test

We checked the performance of PhyloProfile with increasing data size.

In brief, the time required for both importing and plotting the full data (Figure 1), and RAM usage (Figure 2) scales linearly with the size of the data matrix. Plotting of the first 30 genes (default setting; cf point 2. below) 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.

In detail, we assessed the performance of PhyloProfile on a locally installed version using a Macbook Pro CPU core i7 2.8ghz, 8gb ram. 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 (Fig. 1a) or fewer taxa (Fig. 1b), and measured the time to upload and plot the data.

Discussion

(Runtime benchmarking)

Conclusion

Beside the presence/absence pattern of genes across species, PhyloProfile is able to display two additional layers of information. In particularly, the tool facilitates the visualization and exploration of phylogenetic profiles together with the protein feature architectures in an interactive and effective way. Implemented with the dynamic analysis functions, PhyloProfile can offer a reliable analysis of complex phylogenetic profiles.

Outlook (add to the general outlook?)

Limitation for huge data size (runtime)

More useful info (find the time point where genes got duplicated)

Automatic pipeline from sequence to pathway / tree

Distribution analysis of the microsporidian LCA proteins

Introduction

In this analysis, we used PhyloProfile to explore the phylogenetic profiles of the microsporidian LCA proteins. Our aim was to estimate the evolutionary age for those LCA genes to gain more insight about the reduction and expansion history of microsporidia genomes.

Methods

Orthology prediction

For a sufficient data, we use 491 taxa across the tree of life including eukaryote, archaea and bacteria, which are grouped into 44 super taxa as can be seen in the schematic species tree in Figure A‑6. The full list of the taxa under this analysis is described in Table A‑4. We used HaMStR to search orthologs for 1605 microsporidian LCA proteins were searched in the 480 non-microsporidia species using HaMStR approach. To increase the stringency of the ortholog prediction, only sequences that are orthologous with all microsporidia proteins in the current ortholog group can be added into that group. We used only the best 10 HMM hits for performing the reverse BLAST searches. For every search species, only one representative ortholog was selected. Furthermore, we also checked for the co-orthologs during the reverse BLAST search as discussed in Chapter 1/Methods (add link).

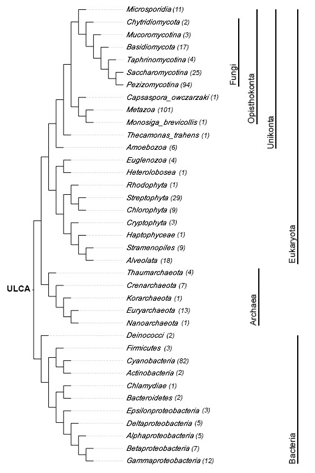


Figure A‑6: (DRAW AGAIN) A cladogram depicts a species tree for all taxa used in the distribution analysis of microsporidian LCA proteins. The number in parenthesis next to the taxon names denotes the number of species in each supertaxon.

Feature architecture similarity score calculation

For a comprehensive analysis and to complement the orthology assignment, we calculated the feature architecture similarity (Koestler *et al.*, 2010) scores (FAS scores) for all pairwise proteins between microsporidia seed and non-microsporidia ortholog. Feature architecture of a protein is the arrangement of different types of protein domains such as PFAM (Finn *et al.*, 2014) or SMART (Letunic *et al.*, 2012) domains, transmembrane domains, low complexity regions, or secondary structures. Comparison of feature architecture between two proteins gives a FAS score between 0 and 1. The higher the FAS score, the more similar those 2 proteins are in term of functional equivalence.

Phylogenetic profile analysis

We applied PhyloProfile 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.

Results

The fast evolving of microsporidia proteins could fail the orthology prediction, especially in the distantly related species. Therefore we additionally used FAS scores as a confidence value for the orthology assignment.



Figure A‑7: (INCREASE FONT SIZE) The distribution of FAS scores for all orthologs of 1605 microsporidian LCA proteins.

The FAS score frequency in Figure A‑7 revealed the 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 and the mean FAS score of in total is up to 0.96.

We clustered 1607 phylogenetic profile of the microsporidia LCA proteins and display the whole profile plot to have an overview about their distribution. Figure A‑8 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.

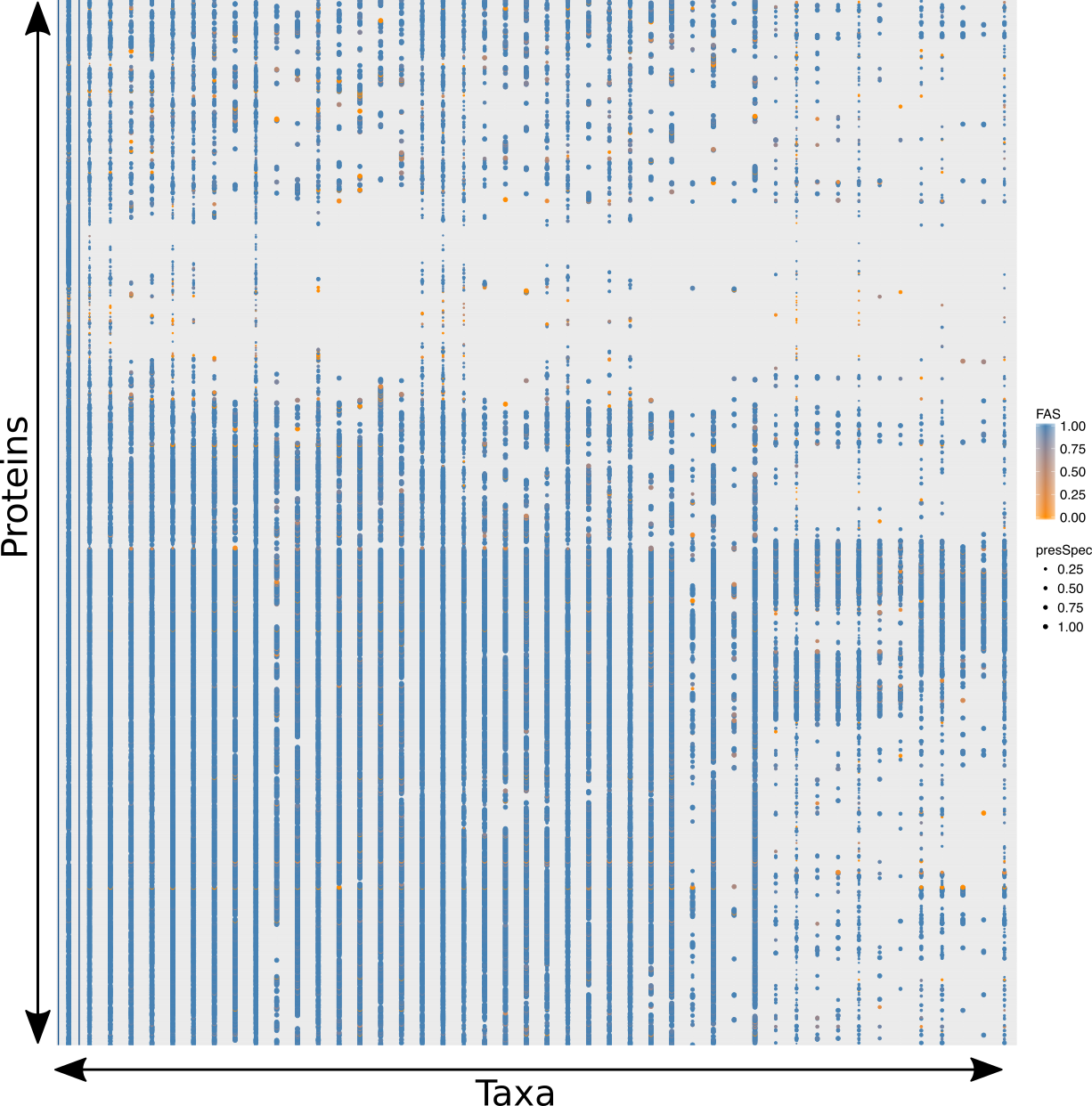


Figure A‑8: The full phylogenetic profile of 1605 microsporidian LCA protein across 491 taxa grouped in phylum level. The color of the points denotes the FAS score between microsporidia and non-microsporidia protein. The size of the points depicts the percentage of species that have orthologs in each phylum.

We then estimated the evolutionary ages for the microsporidian LCA proteins. In accord with the result in Figure A‑8, 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 A‑9).



Figure A‑9: 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.

To investigate the functionality of the 42 microsporidia specific proteins, we used BlastKOALA (Kanehisa *et al.*, 2016) to annotate KEGG Orthology identifiers (Kanehisa *et al.*, 2016) for those proteins. Only 7 of them were linked to KO identifiers (Table A‑1).

Table A‑1: KO annotation for 42 microsporidia specific proteins using BlastKOALA

|  |  |  |
| --- | --- | --- |
| LCA protein | KO identifier | Description |
| OG\_1087 | K17866 | Diphthamide biosynthesis protein 2 |
| OG\_1378 | K09485 | Heat shock protein 110kDa |
| OG\_1378 | K09489 | Heat shock 70kDa protein 4 |
| OG\_1515 | K08803 | Death-associated protein kinase |
| OG\_1710 | K14848 | Ribosome assembly protein RRB1 |
| OG\_2013 | K02155 | V-type H+-transporting ATPase 16kDa proteolipid subunit |
| OG\_2250 | K02896 | Large subunit ribosomal protein L24e |
| OG\_2280 | K02180 | Cell cycle arrest protein BUB3 |

Beside KO annotation, were classify the the microsporidia specific proteins based on Gene Ontology terms (Ashburner *et al.*, 2000) using by Blast2GO v5.0.13 (Götz *et al.*, 2008). Additionally, 12 other microsporidian LCA genes were annotated by GO terms (Table A‑2).

Table A‑2: GO term annotation for 42 microsporidia specific proteins using Blast2GO. All three GO categories were taken into account, in which P is Biological process, C is Cellular component and F is molecular function. (DRAW A TREEMAP AND MOVE THIS TABLE TO APPENDIX)

|  |  |  |
| --- | --- | --- |
| 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 |

Discussion

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 be affected much by applying different FAS cutoffs (see Table A‑3).

Table A‑3: Estimated microsporidia specific proteins by applying different FAS cutoffs.

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

Conclusion

Not surprisingly, due to the compact genomes of the extant microsporidia taxa, most of the proteins in the microsporidian LCA were evolutionary old. Even with a much larger dataset including taxa from all three domains, our result was consistent with the outcome of (Nakjang *et al.*, 2013). About 94% the proteins can be traced to the LCA of all eukaryotes and the other 3% share the same common ancestor with fungal clade. As reported by (Nakjang *et al.*, 2013), they are essential genes for maintenance the survival and development of the microsporidia cells.

Likewise, there are 3% the 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).

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

Introduction

Function assignment is one of the crucial steps in all sequencing projects to characterize the predicted genes or proteins in the new genomes (Gabaldón and Huynen, 2004). There are two main kinds of *in-silico* methods for functional annotation, including structure-based and sequence-based approaches. Because protein structures evolve exponentially slower than their amino acid sequences (Chothia and Lesk, 1986; Williams and Lovell, 2009), the structure-based annotation predictors have advantage over the other, when the sequence similarity is not sufficient enough for retaining the orthologous relationship between two proteins (Adams *et al.*, 2007). Nonetheless, the protein structure prediction process is time consuming and complicated (Baker, 2001). Thus, the sequence-based methods are still commonly used for functional annotation (Loewenstein *et al.*, 2009). Nowadays, with the rapid increase of the newly identified sequences, these methods become more robust since they can be used to automatically annotate a large number of uncharacterized proteins (Sael *et al.*, 2012).

Proteins that are orthologous to each other are likely to have similar functions or retain the key properties (Gabaldón and Koonin, 2013). The quality of orthology-based annotation transfer methods therefore depends strongly on the accuracy of the ortholog prediction. However, orthology inference is indeed a challenging process that is not error-free (Altenhoff *et al.*, 2016). Furthermore, orthology does not always guarantee functional equivalence between proteins (Studer and Robinson-Rechavi, 2009). Therefore, it needs to be combined with other evidences like protein domains for a more accurate function prediction, especially for transferring annotations between distantly related species (Apic *et al.*, 2001; Reid *et al.*, 2007).

Hence, we developed HamFAS, a novel annotation approach based on feature-aware orthology inference. We designed HamFAS to transfer the KEGG Orthology identifiers from the Kyoto Encyclopedia of Genes and Genomes ((Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016)) from the already annotated proteins to their unannotated orthologs. KEGG Orthology identifiers (K numbers or KOs) are used to link proteins with their genomic information including sequences and functional annotations; chemical information such as compound, glycan or reaction; and also their biological metabolic pathways (Kanehisa *et al.*, 2014).

Methods

HamFAS approach

HamFAS is a hybrid approach for KO annotation transfer combining both orthology relationship and protein domain architecture similarity. Figure A‑10 demonstrates the workflow of HamFAS.

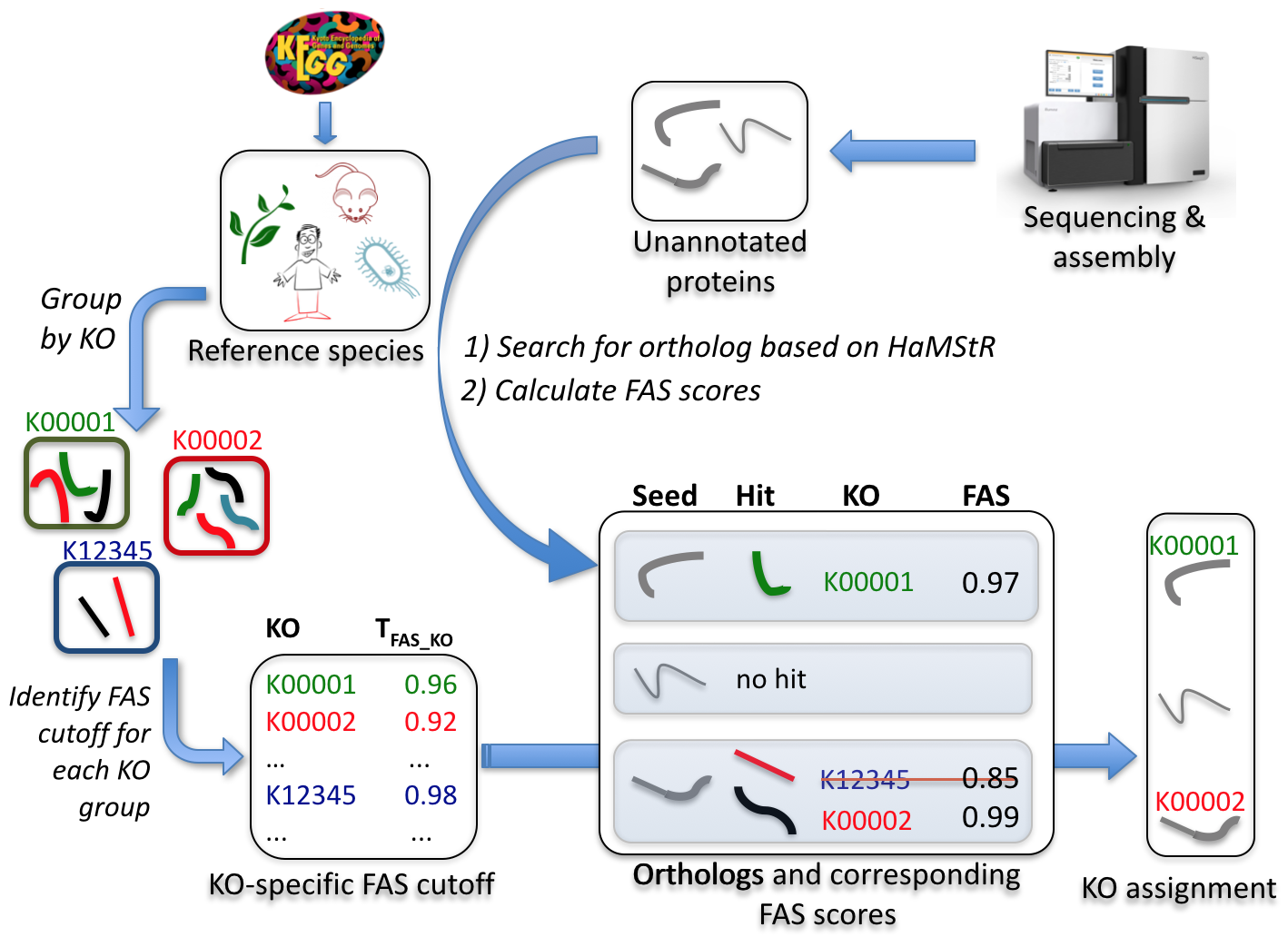


Figure A‑10: KO annotation transfer using HamFAS approach.

First, protein sets of 30 manually KO-annotated reference species (Table A‑5) have been downloaded from KEGG database. We then grouped the obtained sequences by their KEGG Orthology identifiers (KOs). All-vs-all pairwise FAS scores of all reference proteins within a KO group have been calculated. A group's mean FAS score serves then as a KO specific cutoff (TFAS\_KO) that must be exceeded to warrant transfer of its KO identifier to the seed proteins.

Given a list of uncharacterized proteins (seeds), we searched for their orthologs in the reference species using 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. After that, FAS scores between seed proteins and their orthologs were identified. Lastly, we compared the calculated FAS scores with the corresponding TFAS\_KO in order to decide, if the available KEGG identifiers of the paired orthologs can used as potential annotations for seed proteins.

Benchmarking HamFAS

We used *S.cerevisiae* (yeast) as a test species to benchmark the HamFAS approach. The protein set of yeast was obtained from KEGG database. It was divided into two subsets, including one set with 3457 KO-annotated and the second set comprise of 3158 un-annotated sequences. The annotated proteins have been used for evaluating the accuracy of the approach, while the un-annotated set has been used for estimating its sensitivity. The output of HamFAS is also compared with tow state-of-the-art online annotators KAAS (Moriya *et al.*, 2007) and BlastKOALA (Kanehisa *et al.*, 2016) from KEGG.

For this benchmarking purpose, we removed *S.cerevisiae* out of the reference species list to avoid redundant information while performing orthology search. The same reference species have been used for KAAS approach. With BlastKOALA, however, we couldn't remove yeast annotations out of the reference data since this approach use a non-redundant data set established from the whole KEGG's GENES database (Kanehisa *et al.*, 2016) as reference sequences for the KO assignment.

Results

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 A‑11.

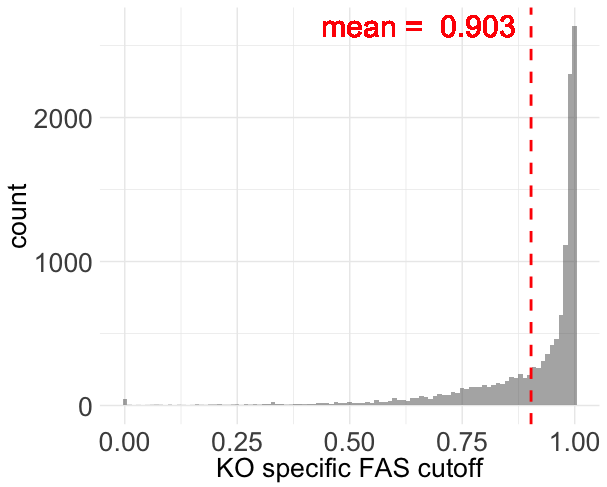


Figure A‑11: (INCREASE X AXIS) Distribution of TFAS\_KO for 12,748 KO groups

Only about 3% of KOs have TFAS\_KO smaller than 0.5, 27% lie between 0.5 and 0.9, while 70% has TFAS\_KO greater than 0.9. The low TFAS\_KO values are caused mostly by the uninformative protein members. Figure A‑12 shows 2 examples for representing a low TFAS\_KO group (K00542) and a high TFAS\_KO group (K0788).

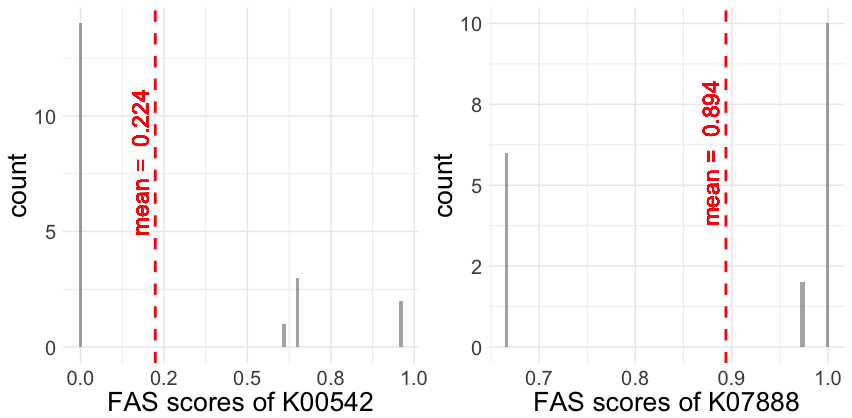


Figure A‑12: FAS score density of KO group K00542 (left) and K07888 (right)

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 TFAS\_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 TFAS\_KO.

Benchmarking result

The specificity of HamFAS approach

The data used for this testing was the KO-annotated yeast proteins. For assessing the specificity, we calculated the recall, precision and F1 score for HamFAS and compared them with the one of BlastKOALA and KAAS.

recall = TP / (TP + FN)

precision = TP / (TP + FP)

F1 = (2\*precision\*recall)/(precision+recall)

Table A‑4 shows the evaluations of HamFAS, BlastKOALA and KAAS. HamFAS performed best in term of precision, while F1-score is lower then KAAS due to its lower recall. Interestingly, the latest annotation tool from KEGG, BlastKOALA, has the lowest scores in both recall and precision.

Table A‑4: Recall, precision and F1-score of HamFAS in comparison to BlastKOALA and KAAS. Second column shows values of HamFAS after filtering the orthology assignment with InParanoid's orthologs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

For gaining more confidence about the ortholog prediction result from HaMStR, we evaluated the annotation transfer again using only orthologs that obtained by HaMStR and further supported by InParanoid (O’Brien *et al.*, 2005; Sonnhammer and Östlund, 2015). Predicted KOs of 188 yeast proteins has been removed after filtering based on InParanoid's orthologs. It resulted in a decrease of the recall and F1-score of HamFAS approach. However, the precision is not affected (see Table A‑4). The FAS scores of the unsupported orthologs are slightly smaller than the ones of the supported orthologs (p-value = ... DO SOME TEST HERE, U-TEST?.), with the mean score of 0,918 and 0,988 respectively (see Figure A‑13).

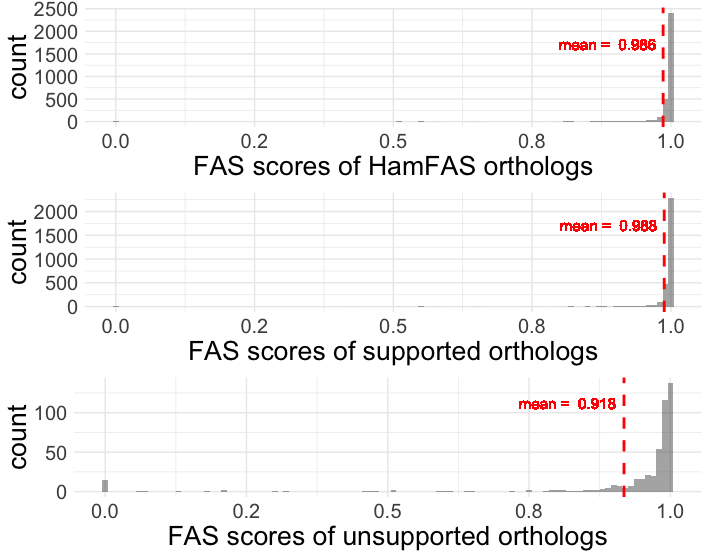


Figure A‑13: FAS score distribution of all HamFAS orthologs, only supported orthologs and unsupported orthologs

We further analyzed the fractions of proteins annotated by HamFAS, BlastKOALA and KAAS. The result is shown in Figure A‑14 with 85,6% of the seed proteins has been annotated by all 3 approaches. While only a small fraction is specific for each tool. In particularly, 2,1% proteins were only annotated by BlastKOALA, and 0,6% in case of HamFAS or KAAS.

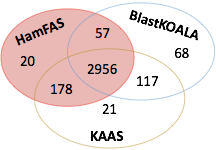


Figure A‑14: Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

For each protein that was annotated by at least two different approaches, we then compared the transferred KEGG identifiers. There is a small difference between the KOs annotated by each approach, which is shown in Table A‑5 below.

Although those KEGG identifiers are different, most of them are "synonymous" KOs. We defined two KOs are "synonymous", if they either have the same EC numbers, same EC classes, same GO numbers, or are the same components in KEGG pathways, responsible for the same reactions.

Some examples of synonymous KOs: (ADD KO NUMBER)

1 KO is very general described (putative ABC transport system ATP-binding protein) while the other is more specific (phospholipid/cholesterol/gamma-HCH transport system ATP-binding protein).

2 KOs have synonym/alternative name: "septin" is synonym with "sporulation-regulated protein 3"; or "tristetraprolin" (ZFP36) and "butyrate response factor 1" (ZFP36L1) are the same.

2 KOs are involved in the same reaction or process: "cleavage stimulation factor subunit 2" and "polyadenylate-binding protein 2" are involved in 3-end formation of pre-mRNAs.

Table A‑5: Compare KEGG identifiers annotated by HamFAS, BlastKOALA and KAAS. Numbers in parentheses are the different KOs after filtered by synonymous KOs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

The sensitivity of HamFAS approach

We used 3158 un-annotated yeast proteins for approaching the sensitivity of HamFAS as well as compared it to the one of BlastKOALA and KAAS. As been seen from Figure A‑15, HamFAS could annotate 257 proteins, in which 164 proteins are HamFAS specific. In comparison to KAAS and BlastKOALA, which annotated 150 and 116 proteins respectively, HamFAS achieved a higher sensitivity among all approaches.

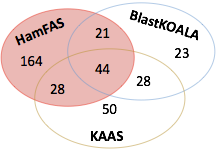


Figure A‑15: Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

Discussion

The specificity of HamFAS

As we have seen from the analysis of the KO-annotation yeast protein set, HamFAS yielded the best precision regardless the supported or non-supported orthology assignment by InParanoid. It indicates the reliability of the annotation transfer result of HamFAS.

The sensitivity of HamFAS

Beside the highest specificity, HamFAS was shown to be more sensitive than BlastKOALA and KAAS with the highest number of proteins that could be annotated. To assess the quality of the proteins that were annotated by HamFAS, we performed some further analyses.

~~First, we filtered the orthology assignment of HamFAS by keeping only orthologs that were also predicted by InParanoid. The number of annotated proteins was reduced from 257 to 150 proteins, in which 55 of them were only annotated by HamFAS approach. Here the~~

Firstly, we check for the length and the domain annotation of proteins, which were annotated only by HamFAS (HamFAS-only proteins) and compared with those of other proteins that were additionally annotated by BlastKOALA and/or KAAS.

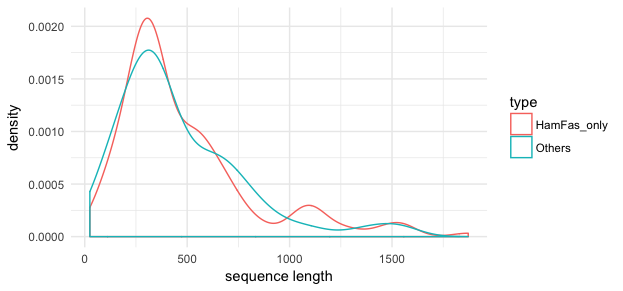


Figure A‑16: Length distribution of HamFAS-only proteins and others

Figure A‑16 and Figure A‑17 show no clear difference between those 2 protein sets (AGAIN, DO SOME REAL TEST HERE). HamFAS-only proteins are not either extremely shorter or longer than other proteins. And the annotation transfer result was not driven by the uninformative domain annotation of those proteins, namely one Pfam domain that led to the high FAS score of the HamFAS-only proteins.

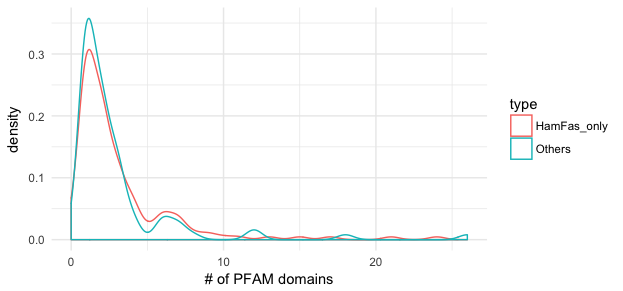


Figure A‑17: Number of Pfam domains distribution of HamFAS-only proteins and others

Additionally, the distribution of FAS scores of all (shouldn't more meaningful when compare HamFAS-only and the others? YES! To keep the data consistent with above analyses) HamFAS orthologs in comparison to HamFAS-only orthologs shown in Figure A‑18 also confirms the comparable similarity (P VALUE) between those two groups of proteins and their orthologs.

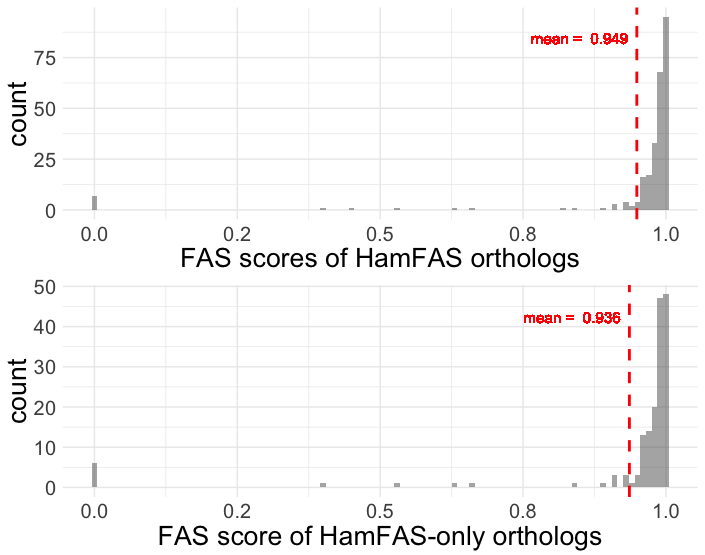


Figure A‑18: FAS score distribution of all HamFAS orthologs and HamFAS-only orthologs

Secondly, we estimated the origin of the annotations for all un-annotated yeast proteins and compared with those of annotated protein set (Figure A‑22).

Figure A‑22: (REDRAW) Origin of KO-annotations for annotated, un-annotated proteins and HamFAS-only proteins of un-annotated set

Unsurprisingly, most annotations of annotated proteins come from their fungal orthologs (75%) while only few of them have obtained annotations from archaea or bacterial taxa (2,4%). In contrary, although a large amount of annotations for un-annotated proteins originated from eukaryotes taxa (78%), there are still 22% (or 27% in case of HamFAS-only proteins) annotations came from distantly related taxa.

Excluding annotations from archaea and bacteria reference orthologs did not affect the accuracy likewise the sensitivity of HamFAS (see Table A‑7 and Figure A‑25). Furthermore, we investigated of the phylogenetic profiles of un-annotated proteins, whose new annotated KOs emerged from non-eukaryotic reference species. This analysis revealed no difference between the HamFAS-only proteins and other proteins that are annotated by both HamFAS and at least one of the tools BlastKOALA and KAAS (see Figure A‑23 and Figure A‑24).

Lastly, we studied the connectivity and KEGG metabolic pathways of the KOs yielded by HamFAS to carry out the biological meaning of those annotations. We calculated the node degree of those proteins in yeast protein-protein-interaction (PPI) networks retrieved from 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).

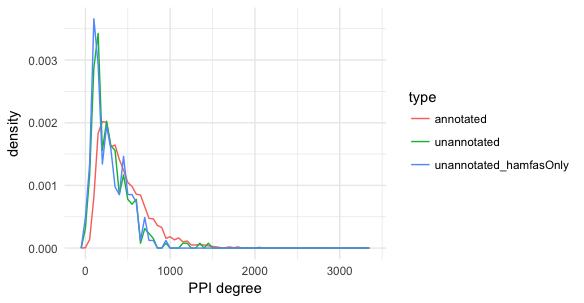


Figure A‑26: The PPI degree distribution of 3 protein sets

Figure A‑26 shows the distribution of PPI degree of the KO-annotated, un-annotated and HamFAS-only proteins within the un-annotated set. The KO-annotated proteins have in general more interacting partners (mean PPI degree 444) than un-annotated and HamFAS-only proteins (mean PPI degree 294 and 275 respectively) (DO STASTISTIC TEST). However, 99% of the proteins of un-annotated set have the PPI degree more than 10, while only 2 proteins don't have any interacting partner.

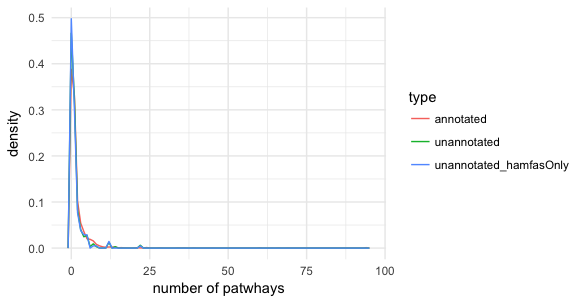


Figure A‑27: Distribution of the number of pathways in which annotated KOs are involved

Beside the PPI degrees, we also determined the number of pathways in which those KOs are involved. All 3 data sets show the same trend in Figure A‑27, namely not less then 50% the KOs belong to at least one KEGG pathway (KO-annotated set 61%, un-annotated set 53% and HamFAS-only protein set 50%) and a large number of KOs were not mapped into any pathway.

A list of pathways that contains the KOs from the HamFAS-only proteins can be found in Figure A‑28. Interestingly, 29 pathways of *S.serevisiae* could be further complemented by new KOs from HamFAS. (See appendix)

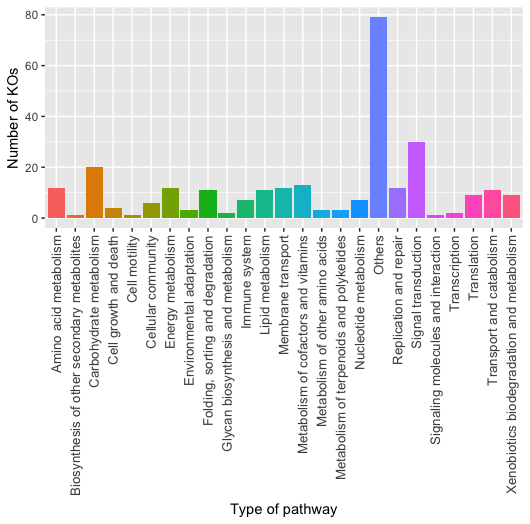


Figure A‑28: The numbers of HamFAS-only KOs distributed into different pathway categories

How different are the phylogenetic profile of KO-annotated proteins and un-annotated protein?

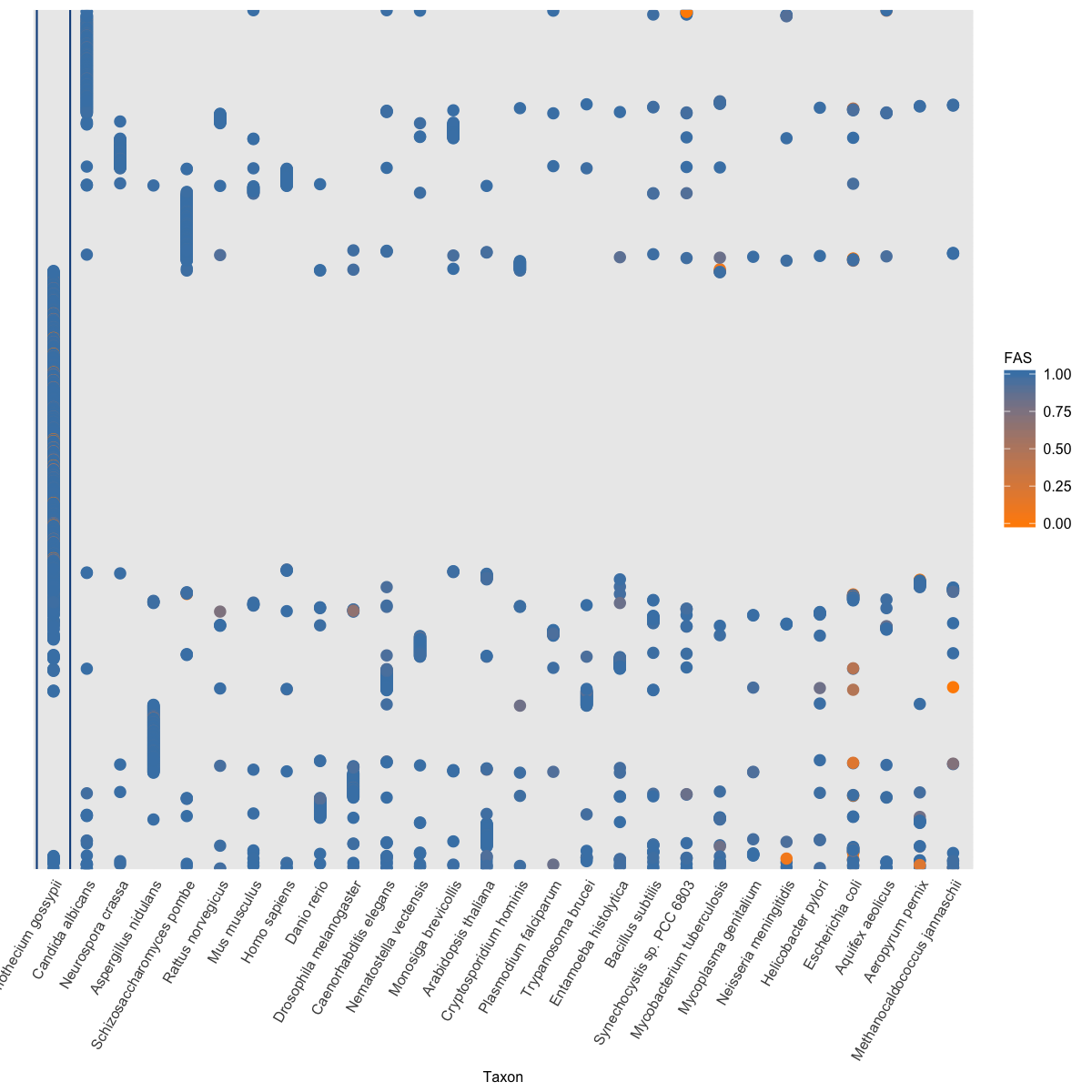


Figure A‑26: Phylogenetic profile of KO-annotated proteins

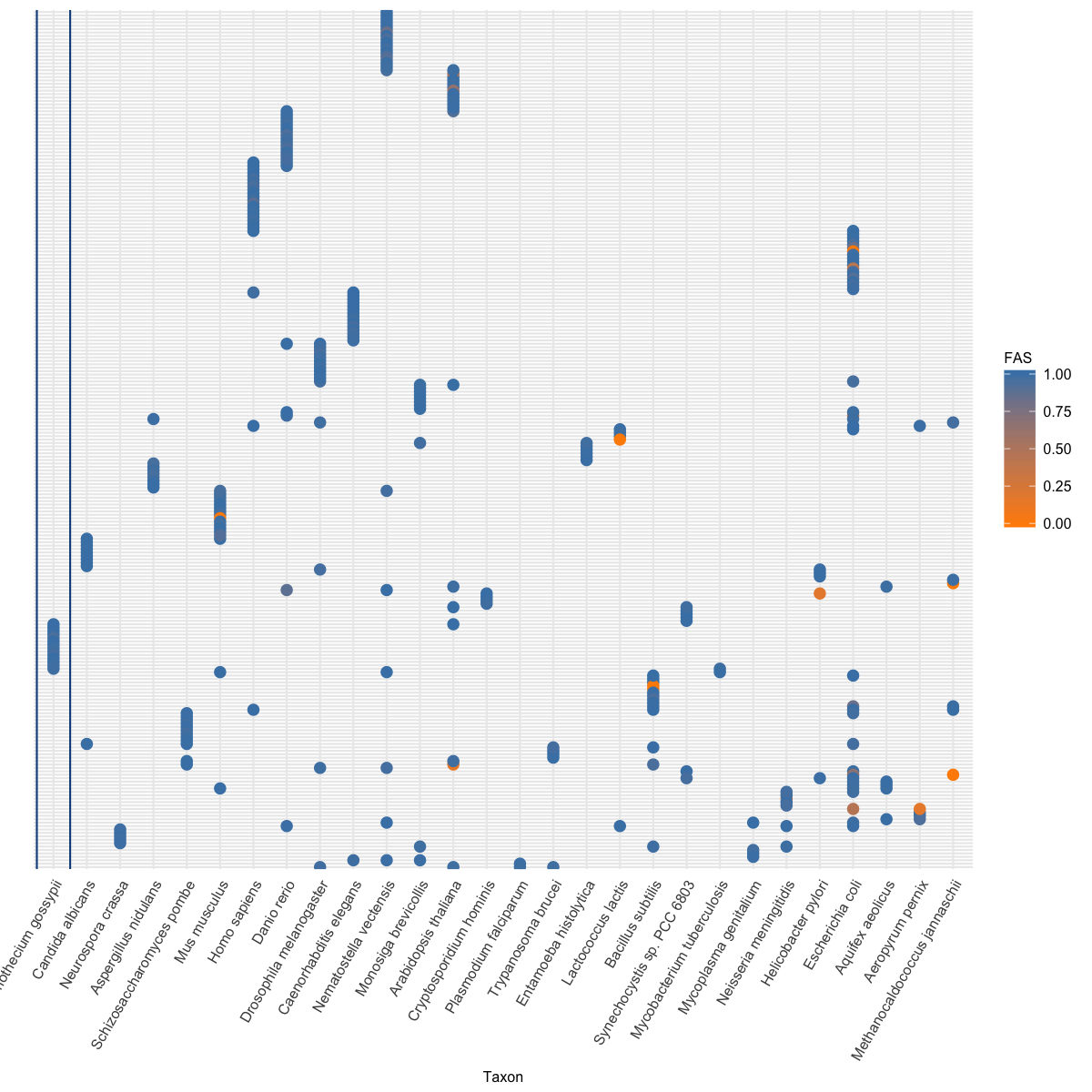


Figure A‑27: Phylogenetic profile of un-annotated proteins

Figure A‑26 and Figure A‑27 show that orthologs of un-annotated proteins are not broadly distributed like the one of annotated proteins. However, most of the proteins in both annotated and un-annotated set have only one ortholog (79% KO-annotated proteins, 80% un-annotated and 80% HamFAS-only proteins. See Figure A‑28). And more than 22% of un-annotated proteins have only orthologs in distantly related reference taxa (more detail in point **Error! Reference source not found.**).

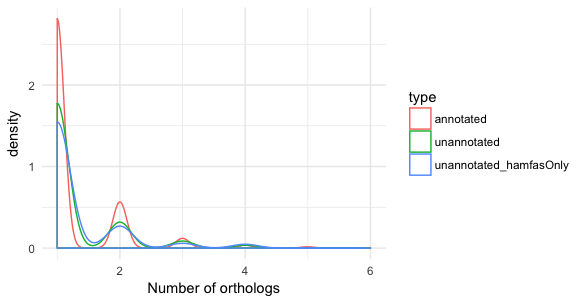


Figure A‑28: Distribution of number of orthologs for KO-annotated, un-annotated and HamFAS-only protein set

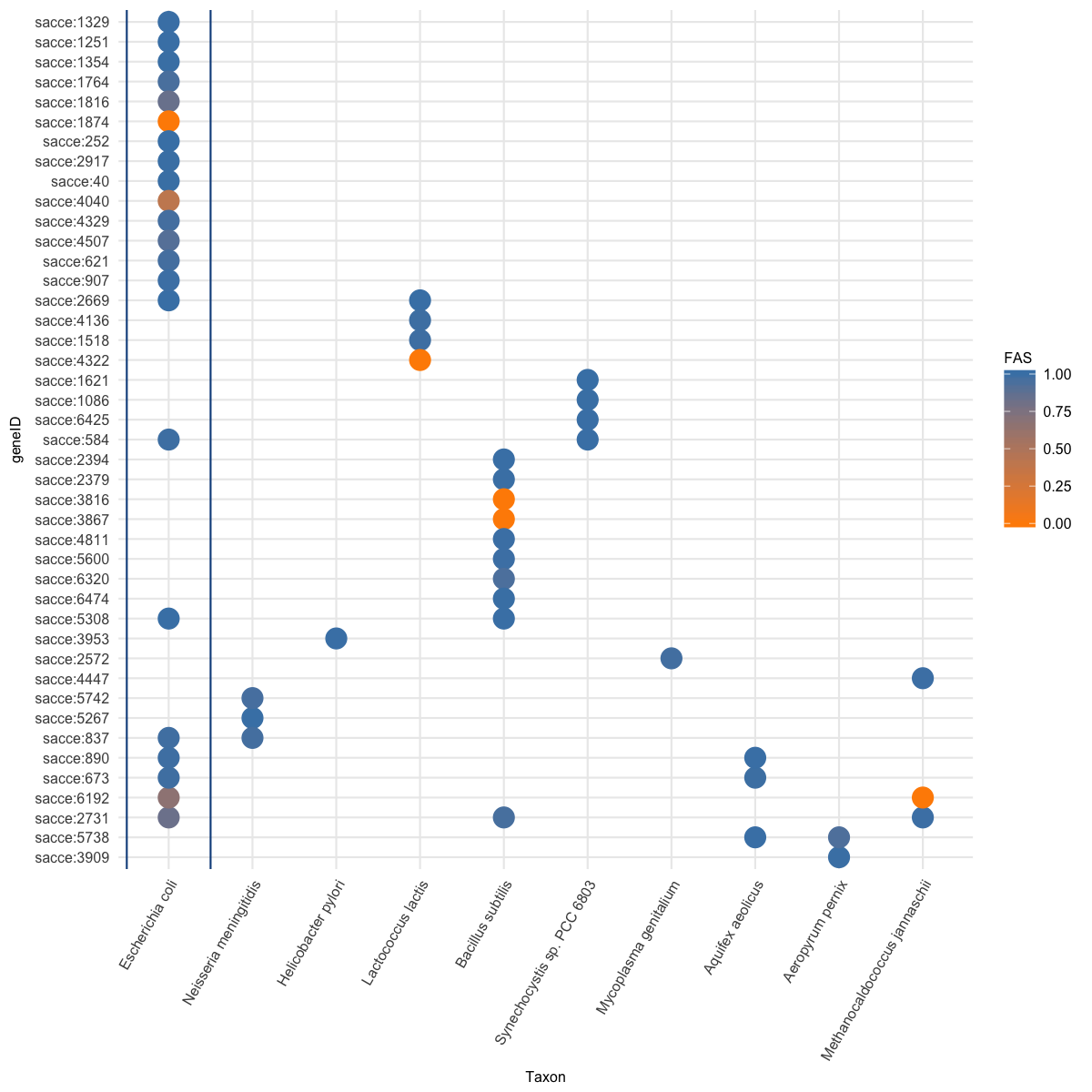


Figure A‑26: Phylogenetic profile of 44 HamFAS-only proteins that annotated based on archaea and bacterial orthologs. (MOVE TO APPENDIX)

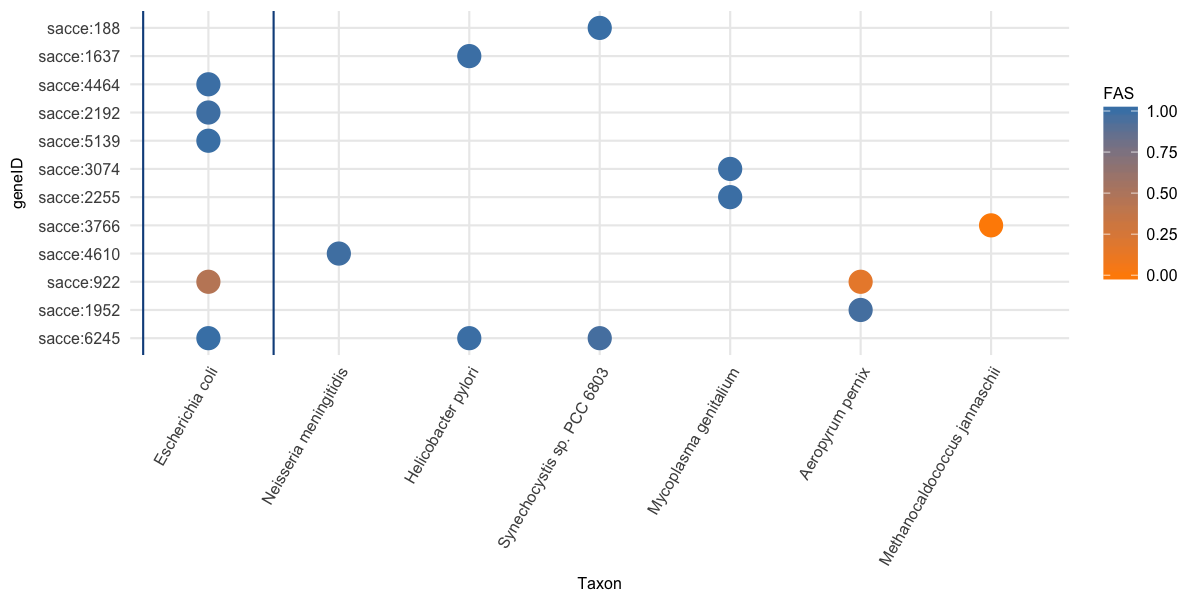


Figure A‑27: Phylogenetic profile of 12 un-annotated proteins that annotated by HamFAS and at least one other approach (BlastKOALA and/or KAAS), where their annotations originate from archaea or bacteria reference taxa. (MOVE TO APPENDIX)

Table A‑6: Recall, precision and F1-score of filtered HamFAS in comparison to HamFAS, BlastKOALA and KAAS by applying on KO-annotated yeast proteins. (MOVE TO APPENDIX)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **HamFAS after filtered** | **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 |

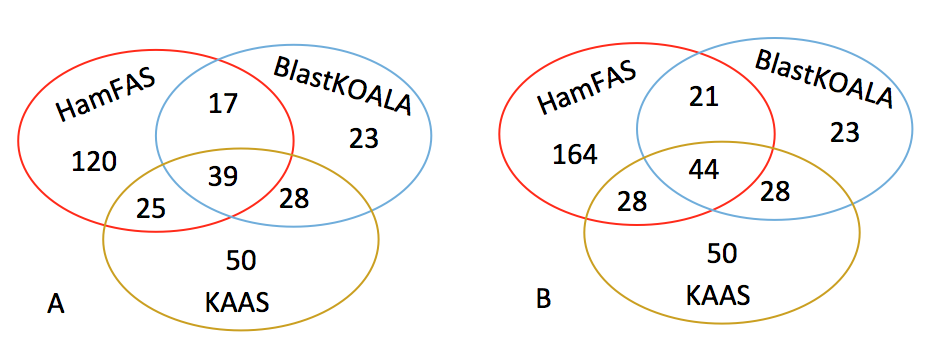


Figure A‑28: Fraction of proteins annotated by BlastKOALA, KAAS and filtered HamFAS (A) or original HamFAS (B) (MOVE TO APPENDIX)

Conclusion

HamFAS appeared to have higher sensitivity and comparable specificity in comparison to the two state-of-the-art annotation tools KAAS and BlastKOALA from KEGG. The difference between the annotation obtained by HamFAS and BlastKOALA/KAAS could be explained by the different orthology procedures were used in each approach. The two KEGG's approaches are based mainly on BLAST with an additional weighting scheme in BlastKOALA (Kanehisa *et al.*, 2016), while HaMStR combines the power of both BLAST and profile HMM search (Ebersberger *et al.*, 2009). On the other hand, HamFAS with the ability of identifying distantly related orthologs using HMM (Madera and Gough, 2002; Alam *et al.*, 2004) could annotate more proteins BlastKOALA or KAAS. The reliability of orthology assignment from HamFAS, or in particularly HaMStR, was confirmed by InParanoid, one of the most accuracy sequence-based orthology inference approaches (REFERENCE).

There was no strong evidence to distinct the difference between HamFAS-only proteins and proteins that are annotated by both HamFAS and other approaches. Their new annotations therefore could be potential candidates for a further experimental analysis.

A feature that makes HamFAS different than BlastKOALA and KAAS is the controllable ability of the annotation process. Users can choose different methods and thresholds to increase or reduce the stringency of the annotation outcome. Besides, HamFAS can be run locally through command lines. It provides a better solution for large-scale analysis than the online tools such as KAAS and BlastKOALA.

The principle factor that affects the annotation result is the accuracy of orthology assignment method. We can increase the stringency of HamFAS by allowing the annotations from only close related species. However, if doing so we will lose the benefit of HaMStR and therefore we have no reason to use HaMStR instead of other more stringent orthology search approaches like OMA or InParanoid. => put to general outlook

Metabolic pathway analysis of the microsporidian LCA proteins

Introduction

Metabolic analysis of microsporidia is still a challenge due to their obligate intracellular growth and short lifetime of its purified spores (Keeling and Fast, 2002). Here we applied HamFAS approach to annotate the microsporidian LCA proteins with the KEGG Orthology identifiers and compared their metabolic pathways with the contemporary species to verify the current hypotheses about microsporidia metabolism and investigate the differences between the metabolism of the microsporidian LCA and the extant species.

Methods

KEGG Orthology annotation

HamFAS approach was used to perform KO annotation for 1605 microsporidian LCA proteins. The reference species is the list of 30 manually curated KO annotated species downloaded from KEGG database (Table A‑5). 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 TFAS\_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; Brocchieri, 2013) 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 (1)

(1)

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 as the lowest normalized distance among all reference proteins that have the matching KO number.

Metabolic pathway analysis

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*, 4/11 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 identifiers) in the pathway and edges are links between those nodes. KO-annotated proteins of each taxon will be then mapped to those reference networks for a connectivity analysis. This connectivity network analysis is implemented into a tool named KEGGcxn.

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

Results

KO annotation for microsporidian LCA proteins

Using HamFAS approach we have annotated 1048 out of 1605 microsporidian LCA proteins with 1344 different KO identifiers.

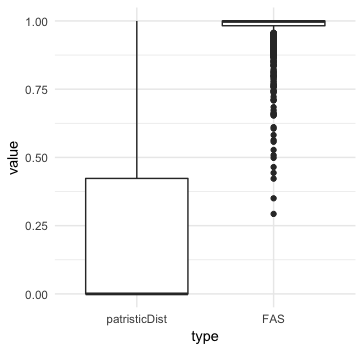


Figure A‑29: Distribution of FAS scores and patristic distances of KO-annotated microsporidian LCA proteins. Blue line represents the conditional mean of FAS score given a patristic distance value.

The distribution shown in Figure A‑29 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 KOs. It indicated that most seed and reference proteins are highly similar in term of domain architectures and a large fraction of the annotations come from the less divergent ortholog sequences to the seed proteins.

The reduced metabolic pathways of microsporidian LCA



Figure A‑30: Pathway enrichment of microsporidian LCA. Colors denote different pathway categories: green for cellular processes, orange for environmental information processing, purple for genetic information processing and pink for metabolism.

Pathway enrichment result of microsporidian LCA is shown in Figure A‑30. While Figure A‑31 shows the fractions of proteins take part in different pathway categories of microsporidian LCA in comparison to other extant species. Microsporidian LCA has more proteins in metabolism than the 4 extant microsporidia species (30% in comparison to 25%, respectively) but still less than the free-living S.cerevisiae (38%).



Figure A‑31: Fractions of proteins distributed in different pathway categories. (NOT NECESSARY)

A more detail of the mapped pathways and number of proteins for each pathway is shown in Figure A‑32. In general, microsporidian LCA has more proteins mapped into pathways in comparison to extent microsporidia species, especially in Cell growth and death, Signal transduction, Folding, sorting and degradation, Carbohydrate and Lipid metabolism. However, it is still very less when compare to S.cerevisiae, a representative of free-living organisms. One possible reason could be, that the number of yeast proteins in this analysis is much higher than the one from microsporidia (3534 yeast proteins versus 950 protein in average for each microsporidia species) <= put into the discussion !!



Figure A‑32: Number of proteins of each taxon (green for microsporidian LCA, orange for E.cuniculi, purple for E.hellem, pink for E.intestinalis, green for N.ceranae and yellow for S.cerevisiae) participates in different KEGG pathways.

We compare the connectivity of annotated proteins between microsporidian LCA and the contemporary species under this study for the core metabolic pathways used in (Nerima *et al.*, 2010) including glycolysis, gluconeogenesis, the Krebs cycle, pentose phosphate pathway, purine and pyrimidine metabolism, and amino acid metabolism.



Figure A‑33: Number of nodes (left) and edges (right) of core pathways for microsporidian LCA, E.cuniculi, E.hellem, E.intestinalis, N.ceranae and S.cerevisiae.

Figure A‑33 shows the comparison between number of nodes and edges in six core metabolic networks for microsporidian LCA and other 5 extant species. The average node degree, average path length and network diameter (the longest shortest paths) can be seen in Figure A‑34. In general, almost all network properties of parasite species are smaller than the free-living species S.cerevisiae, except the path length of Pentose phosphate pathway.



Figure A‑34: Density of average node degree, average path length and diameter (maximal path length) of microsporidian LCA, E.cuniculi, E.hellem, E.intestinali, N.ceranae and S.cerevisiae in 6 core pathways (Glycolysis/Gluconeogenesis, TCA cycle, Pentose phosphate pathway, purine metabolism, pyrimidine metabolism and amino acid metabolism).

Details of network properties for core pathways are shown in Table A‑6.

According to (Fast and Keeling, 2001; Agnew *et al.*, 2003; Keeling and Fast, 2002), microsporidia lacks of mitochondria. But with the presence of genes coding for heat-shock protein 70 (hsp70) in some extant microsporidia species, they suggested that microsporidia ancestor has mitochondria. Those studies also hypothesized that microsporidia will replace pyruvate dehydrogenase complex (PDH) by pyruvate ferredoxin oxidoreductase (PFOR) in order to convert pyruvate into acetyl-CoA and produce NADH. For the common ancestor, we could not find any KOs of 4 PFOR subunits (α, β, γ, δ) in microsporidian LCA, however 3 out of 4 components of PDH were found instead, i.e. pdhA (K00161, OG\_2283, EC=1.2.4.1) and pdhB (K00162, OG\_2084, EC=1.2.4.1) of E1 component, and E3 (DLD) component (K00382, OG\_3281, EC=1.8.1.4) (E2 DLAT, K00627, EC=2.3.1.12 not found). Note that E1 is also be found in N.locustae (Fast and Keeling, 2001) and Encephalitozoon (Katinka *et al.*, 2001). Figure A‑35 shows the mapped microsporidian LCA proteins into the reaction converting pyruvate into acetyl-CoA.

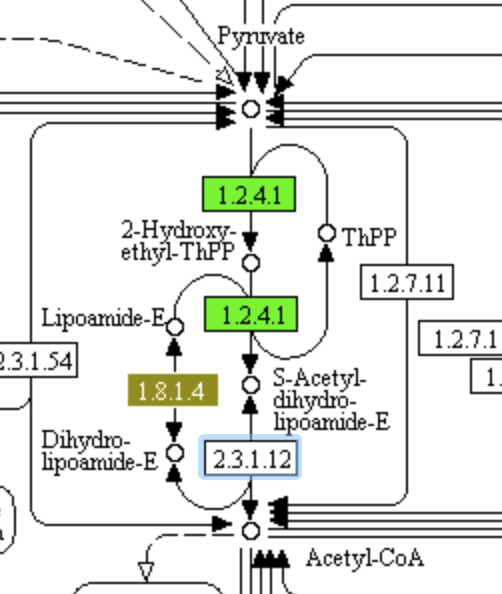


Figure A‑35: (NEED TO BE REDRAWN) The reaction Pyruvate -> Acetyl-CoA with help of pyruvate dehydrogenase complex (PDC). Proteins present in microsporidian LCA are highlighted.

The role of E1 component and the reason for the presence of hsp70 genes in extant microsporidia is still unclear (Together these genes (alpha and beta from E1) provide the first evidence for mitochondrion-derived metabolic activity in microsporidia, but it is still unclear whether they are involved in core energy metabolism in a mitochondrion as in other eukaryotes, or if they have been conscripted into some other pathway during the un- usual course of evolution in microsporidia.) (Fast and Keeling, 2001). But the presence of two subunits of E1 and the component E3 together with the hsp70 proteins (OG\_1157 and OG\_1803 with KO K03283) in the LCA emphasizes the origin hypothesis of mitochondrion in the microsporidia ancestor.

**LOST PATHWAYS**

(7) Yet in microsporidia, endoparasitic fungi living at the limits of cellular streamlining, oxidative phosphorylation has been lost: energy is obtained directly from the host or, during the dispersive spore stage, via glycolysis. It was therefore surprising when the first sequenced genome from the Enterocytozoonidae – a major family of human and animal-infecting microsporidians – appeared to have lost genes for glycolysis. (Wiredu Boakye *et al.*, 2017) See fig 1 and 2 as an example how to represent the presence/absence genes in pathways

Despite the hint about the presence of mitochondrion, it still supposes that microsporidian LCA also lack of TCA cycle, electron transport chain and oxidative phosphorylation pathway like the extant species and other amitochonriate species (Keeling, 2009; Keeling and Fast, 2002; Wiredu Boakye *et al.*, 2017). 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.

Microsporidia mostly uptake ATP from 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 found putative major facilitator superfamily (MFS) transporters in the microsporidia T.hominis. We searched for the transport proteins in the microsporidian LCA and found two MFS transporter and 6 ATP-binding cassette (ABC) transporters (see Table A‑7).

Table A‑8: Microsporidian LCA MFS and ABC transporters.

|  |  |  |
| --- | --- | --- |
| LCA protein | KO identifier | 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 |

Beside up taking ATP using ATP transporters, microsporidia is well-known that they use glycolysis to produce ATP (Keeling and Corradi, 2011; Keeling and Fast, 2002; Heinz *et al.*, 2012).

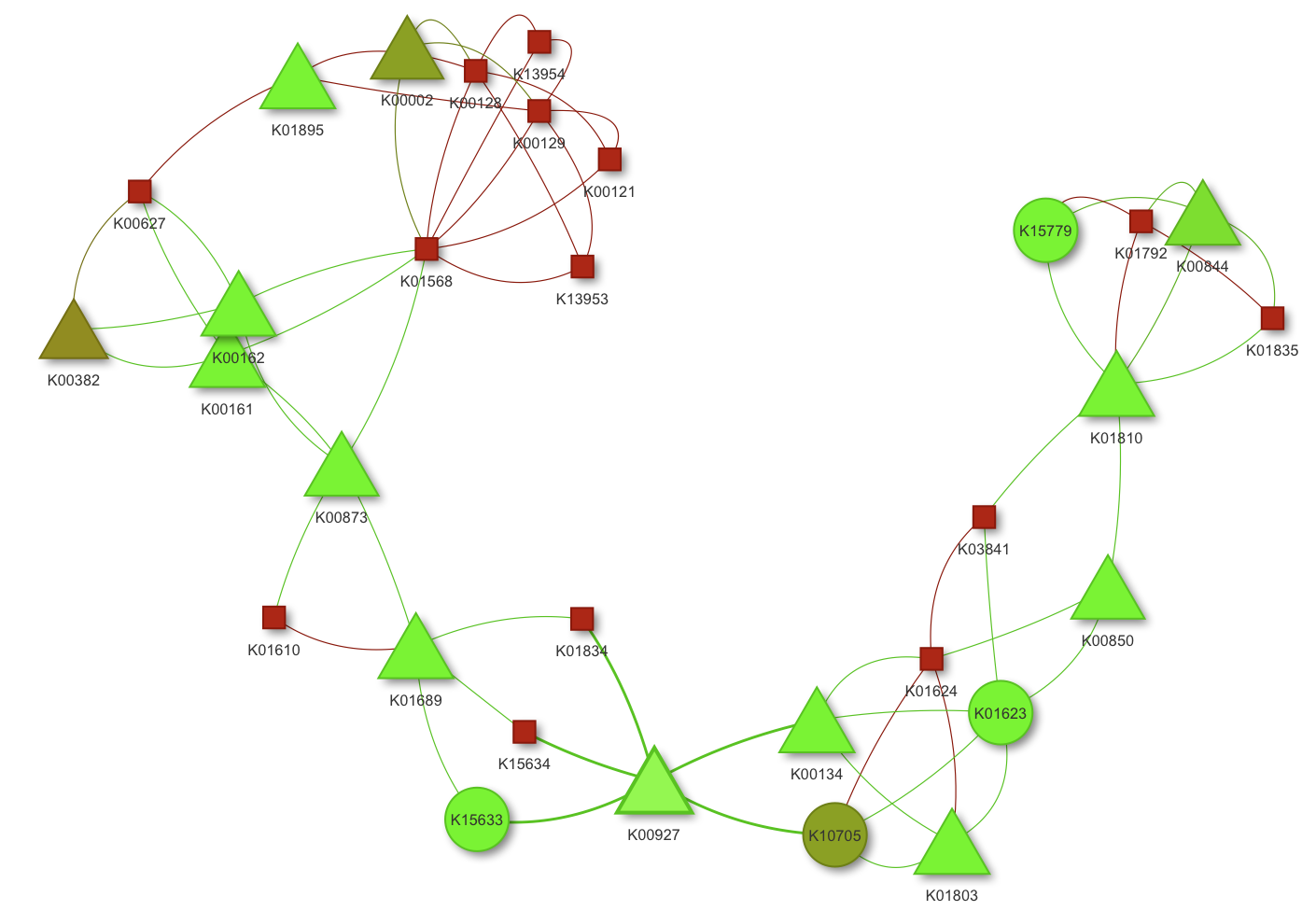


Figure A‑36: Connectivity network of annotated proteins in glycolysis/gluconeogenesis pathway. Circles are microsporidian LCA only proteins. Squares are S.cerevisiae only proteins. Triangles are mapped proteins from both microsporidian LCA and S.cerevisiae. Color denotes the patristic distance between the microsporidia seed protein and the reference ortholog, where the transferred KO annotated comes from (this color code applied only for microsporidian LCA proteins) (ADD COLOR SCALE)

The connectivity network in Figure A‑36 shows the mapped microsporidian LCA and S.cerevisiae proteins into KEGG glycolysis/gluconeogenesis pathway (unmapped nodes are removed). Except K10705 (dark green circle), all other proteins are also present in contemporary microsporidia species. Although the number of nodes of microsporidia is less than S.cerevisiae, they are highly connected and the diameter of microsporidia network is almost the same as the one of S.cerevisiae (10 vs 11). It indicates that microsporidian LCA has almost all keys enzymes for this pathway and can perform the same reactions yeast, except some reactions in one end of the network. See Figure A‑37 for the reference KEGG glycolysis/gluconeogenesis pathway mapped with microsporidian LCA proteins.

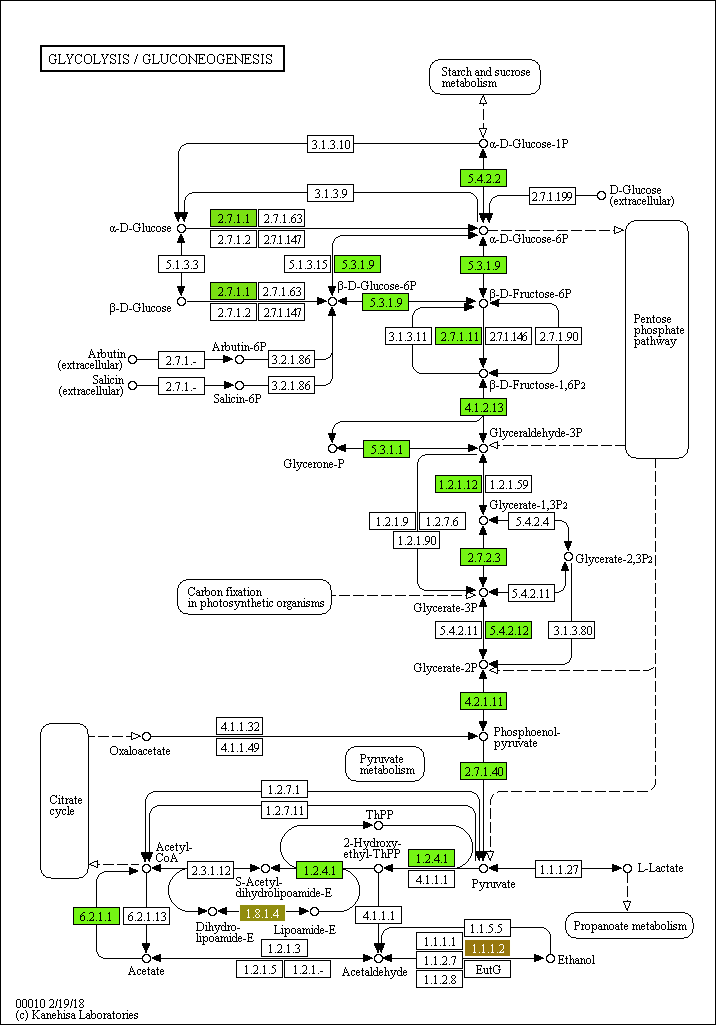


Figure A‑37: Glycolysis/Gluconeogenesis pathway. Annotated microsporidian LCA proteins are highlighted in green. Blank box are unmapped proteins in reference KEGG pathway. (REDRAWN WITH EXTANT MICROS PROTEINS AND MAYBE ALSO YEAST)

According to (Keeling and Fast, 2002; Keeling and Corradi, 2011; Heinz *et al.*, 2012), another core carbon metabolism has been confirmed to be present in microsporidia is the pentose phosphate pathway. This is also true with the LCA, since the pathway mapped with its proteins is the same as the extant microsporidia pathway (Figure A‑38).

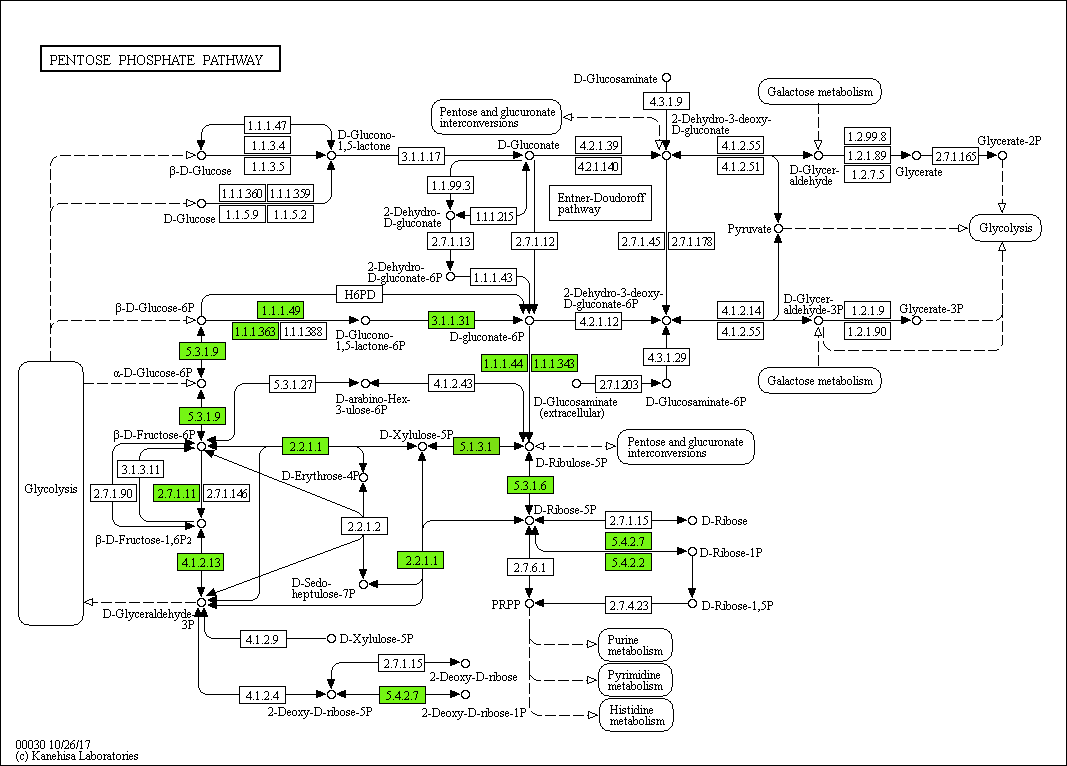


Figure A‑38: Pentose phosphate pathway

The primary carbohydrate storage trehalose is thought to be very essential for the survival and germination of microsporidian spore (Dolgikh *et al.*, 1997; Vandermeer and Gochnauer, 1971; Agnew *et al.*, 2003; Heinz *et al.*, 2012). Enzymes for trehalose synthesis and degradation in extant microsporidia (Keeling and Corradi, 2011; Vandermeer and Gochnauer, 1971; Heinz *et al.*, 2012; Méténier and Vivarès, 2001) have also been found in the LCA including the trehalose 6-phosphate synthase (EC 2.4.1.15 and 2.4.1.347) and alpha-trehalase (EC 3.2.1.28). Figure... demonstrates the starch and sucrose metabolism of microsporidia.

~~The same as contemporary microsporidia species, the LCA also has trehalose 6-phosphate synthase (K00697, EC 2.4.1.15 and 2.4.1.347) and alpha-trehalase (K01194, EC 3.2.1.28), which are involved in the synthesis and degradation of trehalose.~~ (ADD A TABLE WITH ALL ANNOTATED PROTEINS IN THIS ANALYSIS IN THE SUPPLERMENTARY)

DRAW STARCH AND SUCROSE METABOLISM OF MICROSPORIDIA

Having the obligate parasitic life-style, microsporidia tents to uptake nucleotide from the host than produce by themself (Heinz *et al.*, 2012; Dean *et al.*, 2016). Microsporidian LCA lack ribose-phosphate pyrophosphokinase (K00938, EC 2.7.6.1), which converts ribose 5-phosphate into phosphoribosyl pyrophosphate (PRPP) for the de-novo purine and pyrimidine synthesis. IMP cyclohydrolase (K11176, EC 3.5.4.10) making inosine monophosphate IMP and UMP synthetase (K13421, EC 2.4.2.10 & 4.1.1.23) making UMP from PRPP are also not found. The missing of those enzymes indicates that microsporidian LCA is unable to de-novo synthesize both purines and pyrimidines.

See Heinz 2014 for purine and pyrimidine pathway (Fig 1)

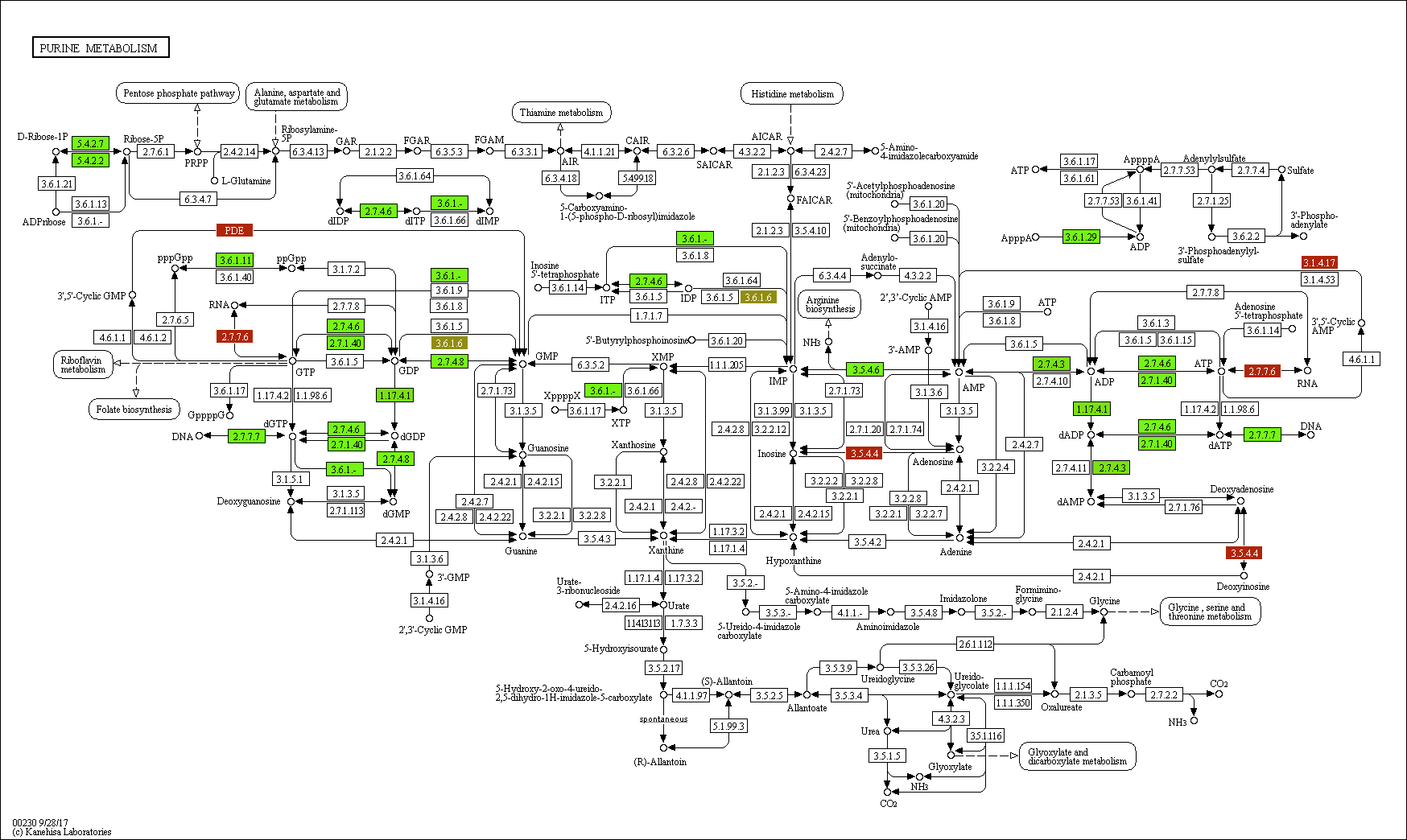


Figure A‑39: Purine

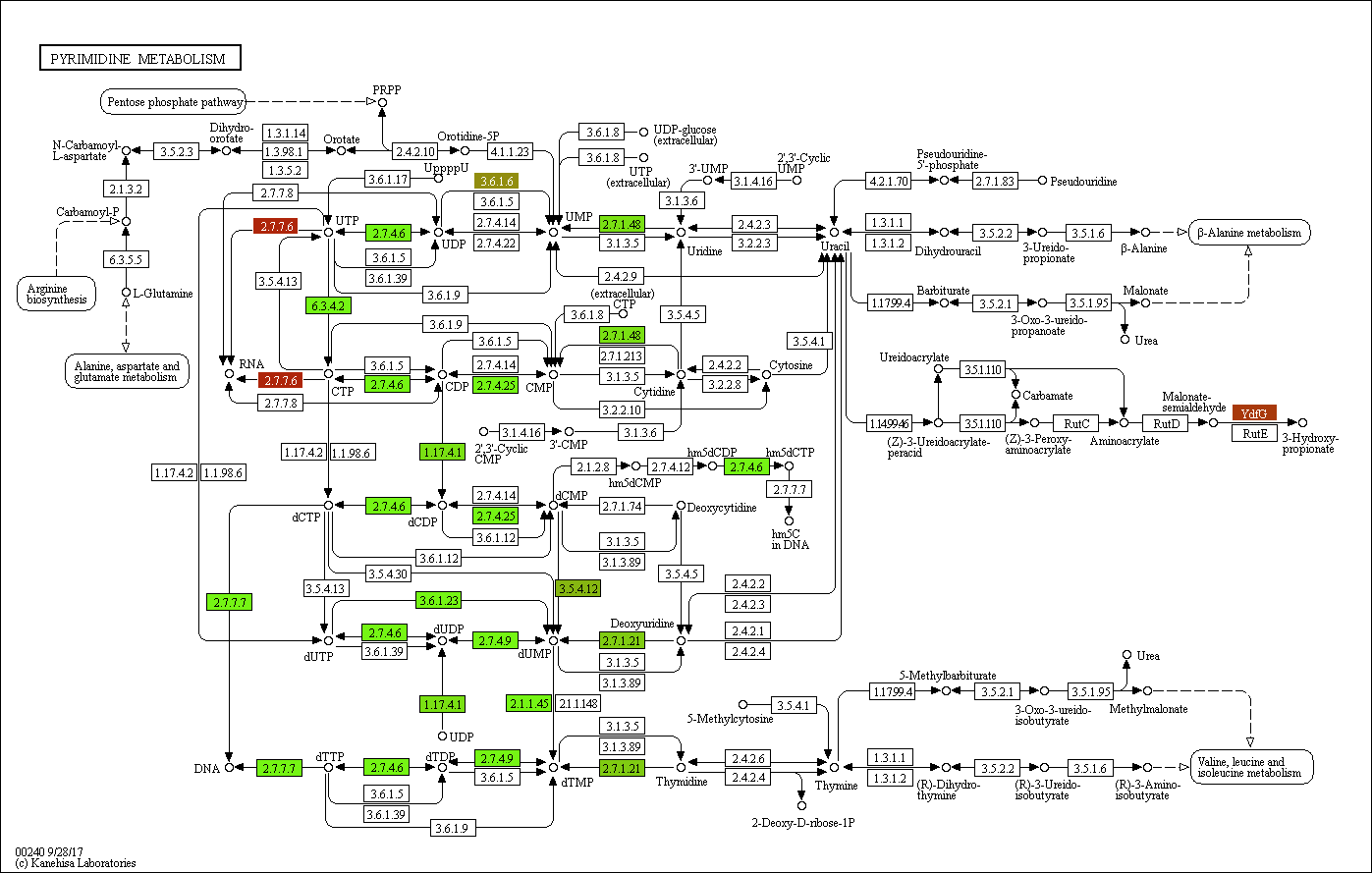


Figure A‑40: Pyrimidine

By that reason, microsporidia need to import nucleotides from the hosts using nucleotide transport (NTT) proteins. KO identifier K03301 of four NTT (NTT1, NTT2, NTT3, NTT4) proteins (Heinz *et al.*, 2014; Dean *et al.*, 2016) have been found in three different microsporidian LCA proteins. Based on studies of (Dean *et al.*, 2016; Tsaousis *et al.*, 2008; Heinz *et al.*, 2014), those NTT proteins are the result of horizontal transfer event from bacteria.



Figure A‑41: Phylogenetic profile of 3 microsporidian LCA NTT proteins

Figure A‑41 shows the phylogenetic profile of 3 microsporidian LCA NTT proteins. All three proteins have bacterial orthologs in Chlamydiae phylum. The domain annotation of a microsporidia protein in comparison with its bacterial ortholog is shown in Figure A‑42. They both contain 11-12 transmembrane domains, as commonly observed in bacterial NTT proteins (Tsaousis *et al.*, 2008; Winkler and Neuhaus, 1999).

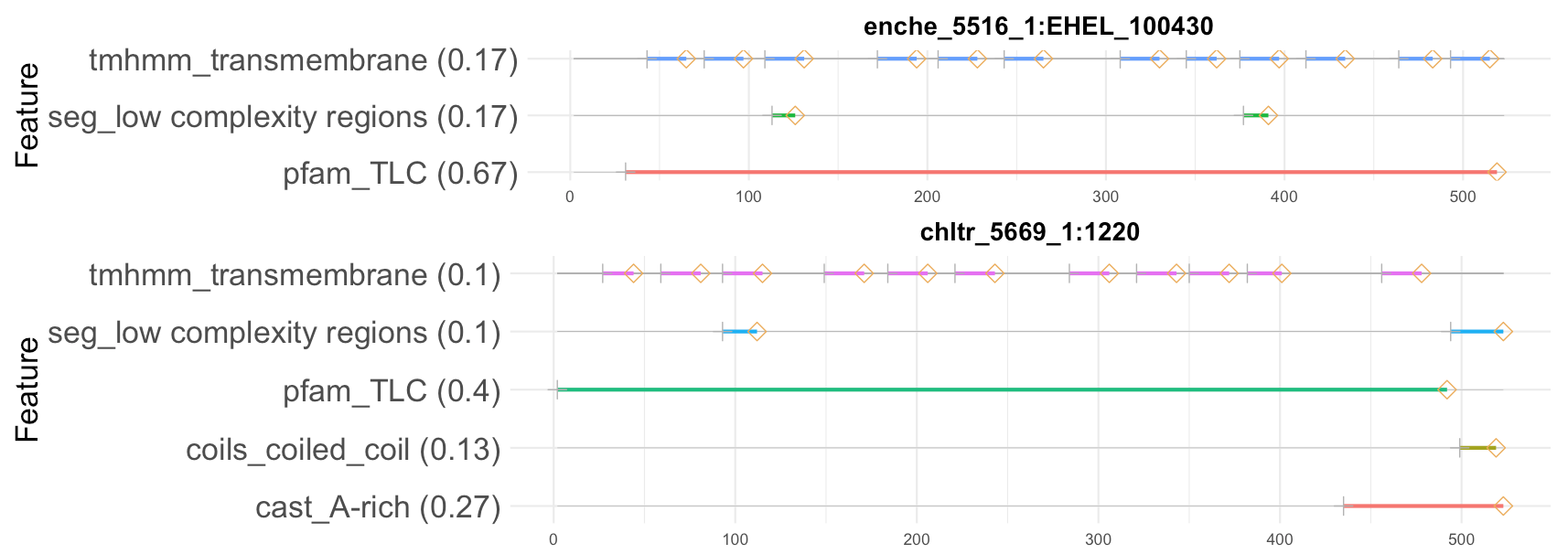


Figure A‑42: Domain architecture of E.hellem protein (enche\_5516\_1:EHEL\_100430) and its ortholog (chltr\_5669\_1:1220) of the bacteria Chlamydia trachomatis. (REMOVE WEIGHT)

4 NTT proteins (

NTT1 - Q8SRA2 - ECU08\_1300 - K03301 <=> OG\_1062,

NTT2 - Q8SUF9 - ECU10\_0540 - K03301 <=> OG\_1062,

NTT3 - Q8SUG0 - ECU10\_0520 - K03301 <=> OG\_3238,

NTT4 - Q8SUG7 - ECU10\_0420 - K03301 <=> OG\_3237 Heinz, Hacker & others 2014)

Check the origin of NTT (also in other species) => see fig 2C (Nakjang 2013)

(9) Genes encoding a fatty acid synthase complex are lacking, which supports the uptake of host-derived fatty acids (El Alaoui *et al.*, 2001; Katinka *et al.*, 2001).

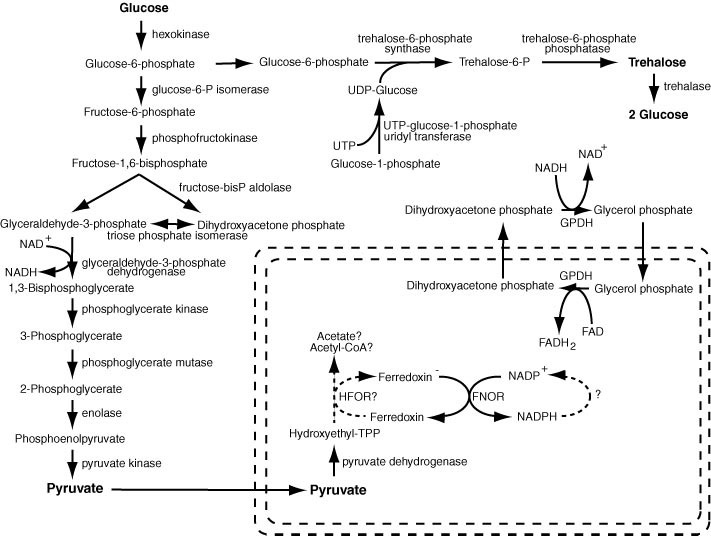


Figure A‑43: carbon pathway (Keeling and Fast, 2002)

Discussion

Conclusion

Our analysis of microsporidian LCA metabolic pathways acquired the consistent results with the finding other studies. Microsporidian LCA, as well as the contemporary species, obligatory depends on the host species for their survival. The presence of transport proteins supplements the lack of some main pathways for producing energy and other important compounds. Trehalose seems to be the main carbohydrate storage for microsporidia since the enzymes for de-novo trehalose synthesis and degradation are also be found in the LCA. However, the reason for the existence of mitochondria in the LCA is still unclear, since the pathways that take place in mitochondria are missing. This analysis demonstrates a novel approach for in-silico studying the metabolic network of microsporidia or any other species.

can replace extant microsporidia by contemporary microsporidia

scheme (used for figure, e.g. scheme of possible energy metabolism in microsporidia)

interpretation

investigation

largely poorly understood

indeed

In essence

illustrating

microsporidian inventions

opisthokont ancestor

# Appendix

Tables

Table A‑1: The representative data set containing eleven extant microsporidia species that was used for the estimation of the microsporidia last common ancestor protein set. The protein sequences of those taxa were downloaded from JGI (Nordberg *et al.*, 2014) and MicrosporidiaDB, where data were obtained from the sequencing project of Broad Institute (Aurrecoechea *et al.*, 2011).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Name | Source | Strain | Number of proteins |
| 1 | *Encephalitozoon hellem* | JGI | ATCC 50504 | 1827 |
| 2 | *Encephalitozoon intestinalis* | Broad Inst | ATCC 50506 | 1657 |
| 3 | *Encephalitozoon cuniculi* | Broad Inst | GB-M1 | 1896 |
| 4 | *Nosema ceranae* | Broad Inst | BRL01 | 2057 |
| 5 | *Enterocytozoon bieneusi* | JGI | H348 | 3312 |
| 6 | *Vittaforma corneae* | Broad Inst | ATCC 50505 | 2243 |
| 7 | *Anncaliia algerae* | Broad Inst | PRA339 | 3576 |
| 8 | *Antonospora locustae* | JGI | HM-2013 | 2191 |
| 9 | *Edhazardia aedis* | Broad Inst | USNM 41457 | 4208 |
| 10 | *Vavraia culicis* subsp. floridensis | Broad Inst |  | 2775 |
| 11 | *Nematocida parisii* | Broad Inst | ERTm1 | 2659 |

Table A‑2: (AGAIN WITH FULL NAME) 24 taxa used for extent the initial homologous groups including 17 non-microsporidia species used in the phylogenetic study of (Capella-Gutiérrez *et al.*, 2012) and other 7 out-group taxa that are written in red.

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Name | No. | Name |
| 1 | *S.pombe* | 14 | *M.brevicollis* |
| 2 | *C.albicans* | 15 | *N.vectensis* |
| 3 | *S.cerevisiae* | 16 | *A.queenslandica* |
| 4 | *N.crassa* | 17 | *C.owczarzaki* |
| 5 | *A.nidulans* | 18 | *T.brucei* |
| 6 | *L.bicolor* | 19 | *N.gruberi* |
| 7 | *P.graminis* | 20 | *A.thaliana* |
| 8 | *M.circinelloides* | 21 | *C.reinharditii* |
| 9 | *R.oryzae* | 22 | *P.sojae* |
| 10 | *P.blakesleeanus* | 23 | *C.hominis* |
| 11 | *B.dendrobatidis* | 24 | *P.falciparum* |
| 12 | *S.punctatus* |  |  |
| 13 | *R.allomyces* |  |  |

Table A‑3: Mean length of orthologous and orphan proteins in 11 microsporidia. We used Wilcoxon-Mann-Whitney U-Test to compare the length of those two protein groups. P-value is less then 0.05 meaning that the length of orthologous proteins are significant different to the one of orphan proteins.

|  |  |  |  |
| --- | --- | --- | --- |
| Taxon | Mean length of orthologous proteins | Mean length of orphans | 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: (AGAIN!!) List of 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 | Monosiga\_brevicollis | unikonta |
| 259 | Capsaspora owczarzaki | 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‑5: List of 30 manually KO-annotated reference taxa

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | Name | No. | Name | No. | Name |
| 1 | *A.gossypii* | 11 | *D.melanogaster* | 21 | *A.pernix* |
| 2 | *S.pombe* | 12 | *C.elegans* | 22 | *E.coli* |
| 3 | *C.albicans* | 13 | *M.brevicollis* | 23 | *N.meningtidis* |
| 4 | *S.cerevisiae* | 14 | *N.vectensis* | 24 | *H.pylori* |
| 5 | *N.crassa* | 15 | *E.histolytica* | 25 | *B.subtilis* |
| 6 | *A.nidulans* | 16 | *T.brucei* | 26 | *L.lactis* |
| 7 | *H.sapiens* | 17 | *A.thaliana* | 27 | *M.genitalium* |
| 8 | *M.musculus* | 18 | *P.falciparum 3D7* | 28 | *M.tuberculosis* |
| 9 | *R.norvegicus* | 19 | *C.hominis* | 29 | *Synechocystis sp.* |
| 10 | *D.rerio* | 20 | *M.jannaschii* | 30 | *A.aeolicus* |

Table A‑6: Network properties of core pathways for microsporidian LCA, 4 extant microsporidia species and *S.cerevisiae*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Pathway | Source | Nodes | Edges | Avg. degree | Max degree | Avg. path length | Max path length (diameter) |
| Glycolysis / Gluconeogenesis | LCA microsporidia | 17 | 19 | 2,24 | 4 | 4,04 | 10 |
| Glycolysis / Gluconeogenesis | E.cuniculi | 14 | 14 | 2,00 | 3 | 3,95 | 9 |
| Glycolysis / Gluconeogenesis | E.hellem | 14 | 14 | 2,00 | 3 | 3,95 | 9 |
| Glycolysis / Gluconeogenesis | E.intestinalis | 14 | 14 | 2,00 | 3 | 3,95 | 9 |
| Glycolysis / Gluconeogenesis | N.ceranae | 13 | 12 | 1,85 | 3 | 3,77 | 9 |
| Glycolysis / Gluconeogenesis | S.cerevisiae | 27 | 48 | 3,56 | 9 | 4,59 | 11 |
| TCA cycle | LCA microsporidia | 3 | 3 | 2,00 | 2 | 1,00 | 1 |
| TCA cycle | E.cuniculi | 2 | 1 | 1,00 | 1 | 1,00 | 1 |
| TCA cycle | E.hellem | 2 | 1 | 1,00 | 1 | 1,00 | 1 |
| TCA cycle | E.intestinalis | 2 | 1 | 1,00 | 1 | 1,00 | 1 |
| TCA cycle | N.ceranae | 2 | 1 | 1,00 | 1 | 1,00 | 1 |
| TCA cycle | S.cerevisiae | 20 | 39 | 3,90 | 5 | 2,53 | 4 |
| Pentose phosphate pathway | LCA microsporidia | 10 | 15 | 3,00 | 6 | 1,93 | 4 |
| Pentose phosphate pathway | E.cuniculi | 10 | 15 | 3,00 | 6 | 1,93 | 4 |
| Pentose phosphate pathway | E.hellem | 10 | 15 | 3,00 | 6 | 1,93 | 4 |
| Pentose phosphate pathway | E.intestinalis | 10 | 15 | 3,00 | 6 | 1,93 | 4 |
| Pentose phosphate pathway | N.ceranae | 9 | 13 | 2,89 | 5 | 1,92 | 4 |
| Pentose phosphate pathway | S.cerevisiae | 15 | 32 | 4,27 | 10 | 2,01 | 4 |
| Purine metabolism | LCA microsporidia | 47 | 131 | 5,57 | 40 | 1,85 | 2 |
| Purine metabolism | E.cuniculi | 40 | 107 | 5,35 | 36 | 1,84 | 2 |
| Purine metabolism | E.hellem | 41 | 107 | 5,22 | 36 | 1,84 | 2 |
| Purine metabolism | E.intestinalis | 41 | 110 | 5,37 | 37 | 1,84 | 2 |
| Purine metabolism | N.ceranae | 30 | 55 | 3,67 | 27 | 1,85 | 2 |
| Purine metabolism | S.cerevisiae | 82 | 310 | 7,56 | 55 | 2,80 | 9 |
| Pyrimidine metabolism | LCA microsporidia | 46 | 85 | 3,70 | 40 | 2,05 | 3 |
| Pyrimidine metabolism | E.cuniculi | 38 | 66 | 3,47 | 35 | 1,99 | 3 |
| Pyrimidine metabolism | E.hellem | 39 | 67 | 3,44 | 35 | 2,08 | 4 |
| Pyrimidine metabolism | E.intestinalis | 40 | 69 | 3,45 | 36 | 2,04 | 4 |
| Pyrimidine metabolism | N.ceranae | 31 | 51 | 3,29 | 29 | 1,88 | 2 |
| Pyrimidine metabolism | S.cerevisiae | 65 | 161 | 4,95 | 51 | 2,61 | 8 |
| Amino acid metabolism | LCA microsporidia | 22 | 11 | 1,00 | 3 | 1,33 | 3 |
| Amino acid metabolism | E.cuniculi | 6 | 1 | 0,33 | 1 | 1,00 | 1 |
| Amino acid metabolism | E.hellem | 8 | 1 | 0,25 | 1 | 1,00 | 1 |
| Amino acid metabolism | E.intestinalis | 7 | 1 | 0,29 | 1 | 1,00 | 1 |
| Amino acid metabolism | N.ceranae | 8 | 2 | 0,50 | 1 | 1,00 | 1 |
| Amino acid metabolism | S.cerevisiae | 146 | 299 | 4,10 | 19 | 5,16 | 13 |

Figures

