Introduction

The microsporidia

Microsporidia is a group of obligate intracellular parasites that infect a broad range of species from vertebrates to invertebrates. They are now classified as sister group of fungi. Microsporidia species have genome sizes from 2,3MB to 19,5MB (Heinz et al. 2012). Most of these parasites have a very small number of genes, from 2000 to 4000 genes, which are thought to be very essential that cannot be lost or they are specific genes of microsporidia. The understanding of the evolution of microsporidia is still unclear and there are many question about the reducing process in their genomes, such as when they lost their genes, if the losing process happened once or many times, if those genes lost forever or just temporarily...

In order to find the answer for those questions, the proteins sets of the microsporidia last common ancestor (LCA) will be inferred using hierarchical orthology inference approach. The distribution of these proteins will be then analyzed to find their origins and to get an overview how common they are in the living organisms. Moreover, we try to investigate the functionality of these microsporidia LCA proteins and identify their roles to understand how they can work together with their hosts' metabolic pathways.

Symbiosis life-style

The estimation of the microsporidia last common ancestor protein set

Introduction

Methods

Starting from 11 extant microsporidia species (Table A‑1), we used OrthoMCL (Li et al. 2003) to search for homologous proteins in those 11 microsporidia species. OrthoMCL performed all-against-all BLASTP comparisons for all input data set and clustered homologous groups using Markov Cluster algorithm MCL (van Dongen 2000).

The initial homologous groups were then extended by using HaMStR (Ebersberger et al. 2009) to search for orthologs in other 24 search taxa (Table A‑2). The Hidden Markov Model (HMM) profiles for the initial homologous groups (seed sequences) crated by HaMStR were used to search in the search taxa. The obtained hits were confirmed by re-BLAST search against the protein sets of seed sequences. We took into account here also the co-orthologs for re-BLAST and we limited the HMM hits up to only the best 10 hits.

We identified a core gene set, where we found orthologs in all taxa and each taxon has exactly one orthologous protein. Using the core gene set, we reconstructed a maximum likelihood species tree for those 35 taxa. Firstly, we used ClustalW to align all orthologous groups of the core gene set. Then we created a super-alignment by concatenating the alignments of all core genes. To exclude the alignment columns that are uninformative, we removed columns that have at least 50% gaps. We used ProtTest (Abascal et al. 2005) to find the best fitting model for the tree reconstruction procedure. With the best model parameters obtained from ProtTest, we used RAxML (Stamatakis 2014) to build the maximum likelihood species tree with 100 bootstrap replicates.

Using the principle of minimum evolution (Edwards 1996), we filtered the orthologous group to obtain the final protein set representing the microsporidia last common ancestor. Those final orthologous groups have to have either (1) at least one ortholog from N.parisii (the earliest branch of the microsporidia clade), or (2) at least two orthologs from microsporidia species different than N.parisii and one or more orthologs from non-microsporidia taxa.

Results

OrthoMCL gave 2904 initial homologous groups for 11 microsporidia protein sets. Out of 2904 extended groups, we found 80 groups, where all 11 microsporidia and 24 non-microsporidia taxa are present and each taxon has one representative ortholog. Those 80 groups serve as out core get set for the species tree reconstruction. The super-alignment after de-gapping has a length from 36.616 amino acids. The best model obtained from ProtTest was LG substitution model (Le & Gascuel 2008), GAMMA distribution G , including proportion of invariable sites estimation I & empirical base frequencies F. The input model parameter for RAxML was PROTGAMMAILGF. The reconstructed species tree with bootstrap support values is shown in the Figure A‑1 below.

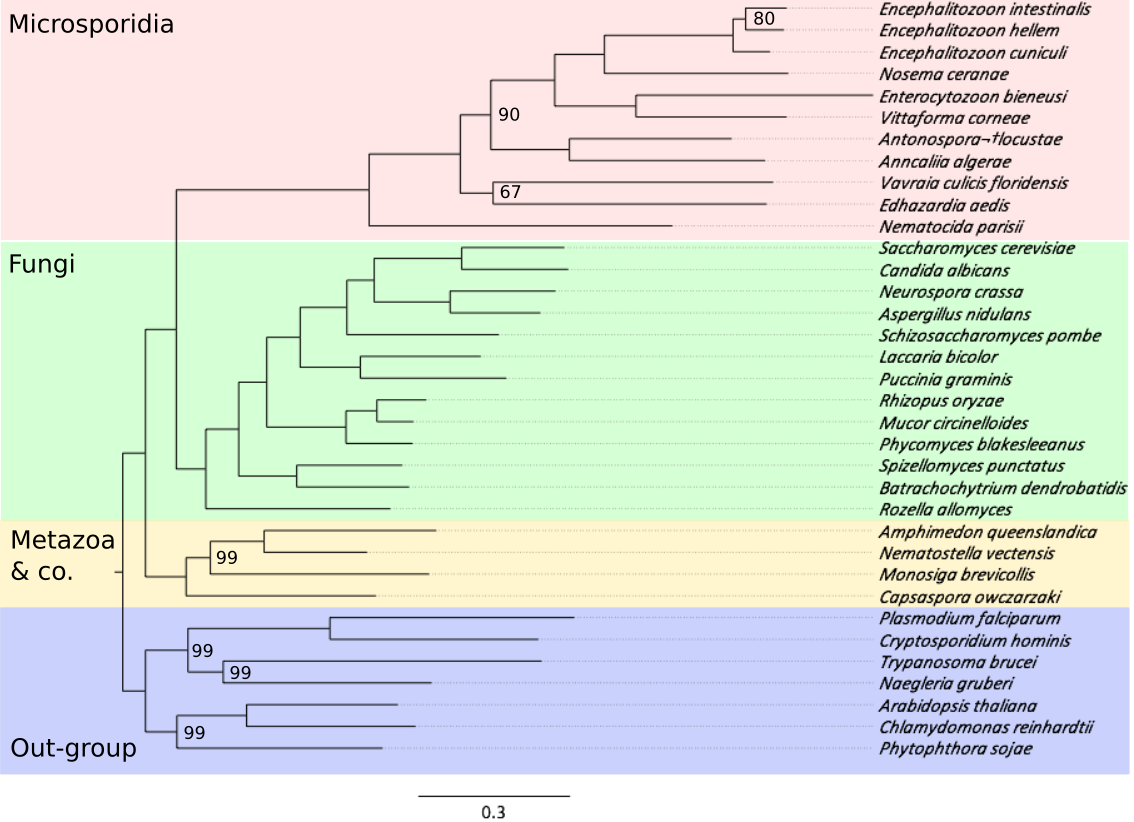


Figure A‑1: Maximum likelihood tree over 35 species. The 11 microsporidia taxa are highlighted in red. Other non-microsporidia taxa include 13 Fungi (green), 2 Metazoa and M.brevicollis, C.owczarzaki (yellow) and 7 out-group species (purple). Node labels denote the bootstrap support and only values 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

Filtered the HaMStR result that did not match the parsimony criteria, we got at the end 1605 final orthologous groups. They present the set of microsporidia last common ancestor proteins.

Discussion

Figure A‑2 shows the fractions of non-orthologous and orthologous proteins in 11 microsporidia species. The Encephalitozoon group is the best example for the compact genome of microsporidia. Where almost 98% of their proteome are orthologous proteins that are shared in other microsporidia species. Only 2% of genes are lineage specific proteins.

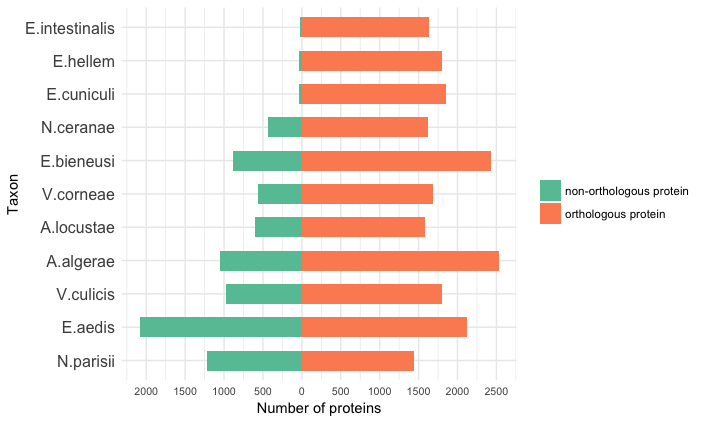


Figure A‑2: Fractions of non-orthologous (orange) and orthologous (green) proteins in different microsporidia species.

Other taxa still have orphan proteins (21% in N.ceranae up to 49% in E.aedis). We have some hypotheses for those orphan proteins.

(1) Wrong gene assignment: length of orphan proteins would be shorter than orthologous proteins.

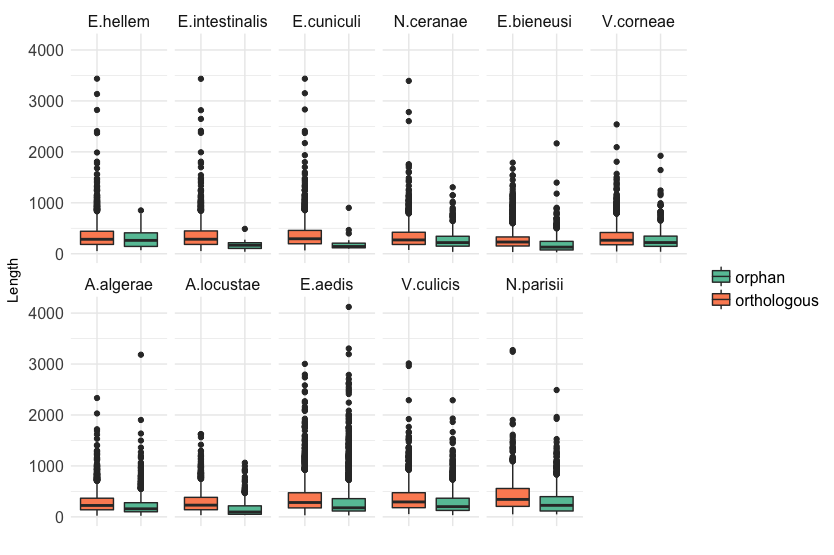


Figure A‑3: Length distribution of orthologous proteins (orange) and orphan proteins (green) in different microsporidia taxa.

Figure A‑3 shows the length distribution of orthologous and orphan proteins in 11 microsporidia taxa. We use Wilcoxon-Mann-Whitney U-Test (is it the same as paired Mann-Whitney?) to compare the two length categories. We found that the lengths of orphan proteins are significantly different (smaller) to the one of orthologous proteins with the significant level of 5%. Only for E.hellem the p-value is 0,20>0,05. But this p-value makes no sense since the number of orphan proteins in E.hellem is too small to make the comparison meaningful.

(2) New invented genes (which have no PFAM annotations), or genes from horizontal gene transfer events (which have new PFAM annotation, which are not found in orthologous proteins), or they cannot be detectable as orthologs (which have the same PFAM annotations as orthologous proteins).

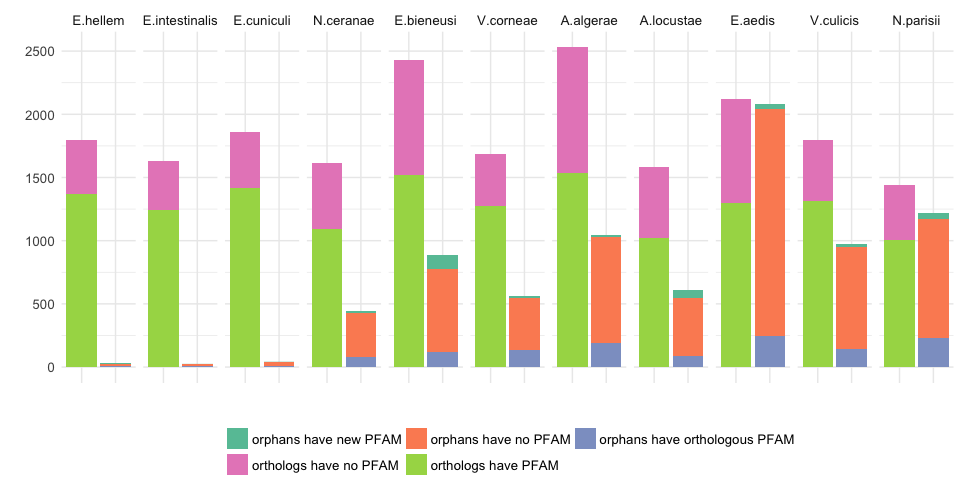


Figure A‑4: Fractions of orthologous and orphan proteins that have and do not have PFAM annotations. The left columns show the number of orthologous proteins that have PFAM annotations (light green) and do not have PFAM annotations (pink). The right columns show the proportion of orphan proteins that 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 microsporidia proteins do not have any PFAM annotation as been shown in Figure A‑4 suggests that most of the orphan proteins are new invented genes after the speciation event that split fungi out of the microsporidia clade.

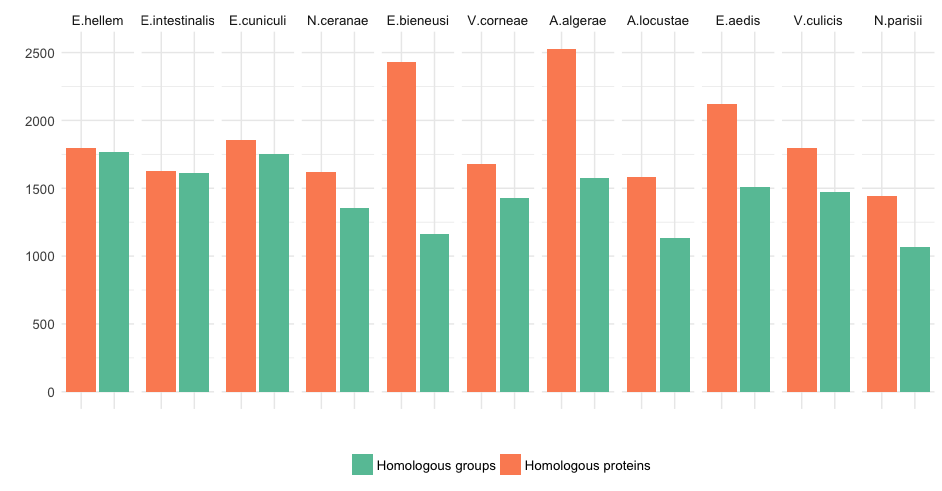


Figure A‑5: The distribution of number of homologous proteins (orange) and number of homologous groups (green) in each microsporidia taxon.

Figure A‑5 shows the relative number of homologous proteins and number of homologous groups in each microsporidia species. In some species, such as E.bieneusi or A.algerae, the number of homologous genes is substantially higher than the number of corresponding homologous groups. We check the number of in-paralogs for each microsporidia taxon in the homologous group. The result in Figure A‑6 shows that there is no evidence for whole genome duplication in any species. But there are some instances where the homologous groups contain more than 10 co-orthologs for one microsporidia species showing the effect of gene dosage (Kondrashov & Koonin 2004).

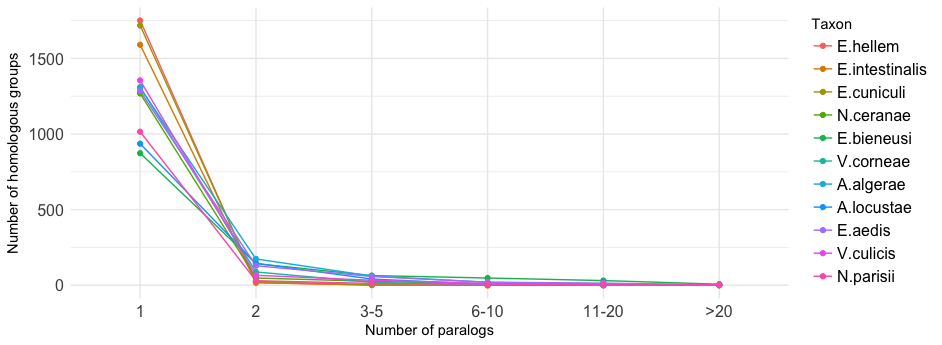


Figure A‑6: The distribution of number of homologous groups as a function of number of in-paralogs. Colors denote different microsporidia taxa.

80 core genes and the origin of microsporidia

The 80 core genes, which are used for reconstructing the species tree in Figure A‑1, shows to be a very good seed set for studying the evolutionary of fungal or even eukaryotic species. It can be used to create very good resolved species trees for a large set of fungal species (ref.) or to investigate the co-evolution of PDI/RhoGID gene clusters (protein disulfide isomerases and Rho guanine-dissociation inhibitors) across the animal phylogeny (Moretti et al. 2017).

More about microsporidia origin (Keeling & Fast 2002; Agnew et al. 2003; Corradi & Keeling 2009) => put in introduction section

The reconstructed species tree shows a stable signal for placing the microsporidia group as the earliest clade of the fungi, which are supported by the recent phylogenetic study of the taxonomy of microsporidia that hypothesis that microsporidia is the sister clade of fungi (Hirt et al. 1999; Capella-Gutiérrez et al. 2012; Katinka et al. 2001; Keeling et al. 2000).

Tree of fungal analysis?

Conclusion

The estimation of microsporidia last common ancestor proteins is the basic step for the whole downstream study. The orthology assignment result agreed with other studies about the fraction of microsporidia only proteins. The length distribution and PFAM annotation analysis indicate that most of the orphan microsporidia proteins are either resulting by wrong gene assignment or they are new invented genes in the microsporidia clade. The good resolved species tree can serve as a fundamental phylogenetic background for filtering the orthology assignment and estimating the set of 1605 proteins for the last common ancestor of microsporidia. This tree supports the hypothesis that microsporidia is a sister group of fungi.

Distribution analysis of the microsporidia last common ancestor proteins

Introduction

To answer the question "How old are the microsporidia last common ancestor proteins"

Methods

We used HaMStR to search orthologs for 1605 microsporidia last common ancestor proteins in 480 taxa across the tree of life including bacteria, archaea and eukaryote, which are grouped into 44 super taxa as you can see in this schematic species tree in the Figure A‑7. The list of all the taxa under this analysis is written in the Table A‑4. The options we used for HaMStR search are -strict, -checkCoorthologsRef, -hit\_limit=10 and -representative.

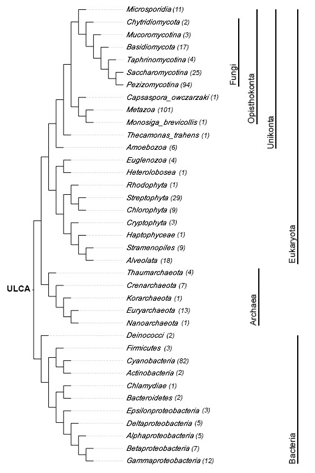


Figure A‑7: A cladogram depicts a species tree for all taxa used in the distribution analysis of microsporidia last common ancestor proteins. The number in parenthesis next to the taxon names denotes the number of species in each supertaxon.

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 pair microsporidia seed proteins and non-microsporidia orthologs. 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, secondary structures, etc. 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.

To visualize the phylogenetic profiles of 1605 microsporidia last common ancestor proteins, we developed a tool named PhyloProfile. This tool is written in R (R Development Core Team 2011) using the Shiny library (https://CRAN.R-project.org/package=shiny). Beside the presence / absence pattern of genes across species, PhyloProfile is able to display two additional layers of information. In particularly, PhyloProfile enables the visualization and exploration of phylogenetic profiles together with the protein feature architectures in an informative and interactive way. Implemented with the dynamic filtering option, PhyloProfile can offer a reliable analysis of phylogenetic profiles with its analysis functions.

We used PhyloProfile to estimate the evolutionary age for microsporidia last common ancestor proteins.

Results

PhyloProfile appears to be vey handy for exploring the informative phylogenetic profile with complementary information.

Almost orthologous proteins have similar feature architectures with microsporidia proteins. It leads to a very high mean FAS score of 0.958 (see the FAS score distribution in Figure A‑8).



Figure A‑8: The distribution of FAS scores for all orthologs of 1605 microsporidia last common ancestor proteins.

Figure A‑9 shows the full phylogenetic profile of 1607 microsporidia last common ancestor proteins across 491 taxa that are grouped into phylum level. A large fraction of microsporidia proteins spread through all studied taxa.

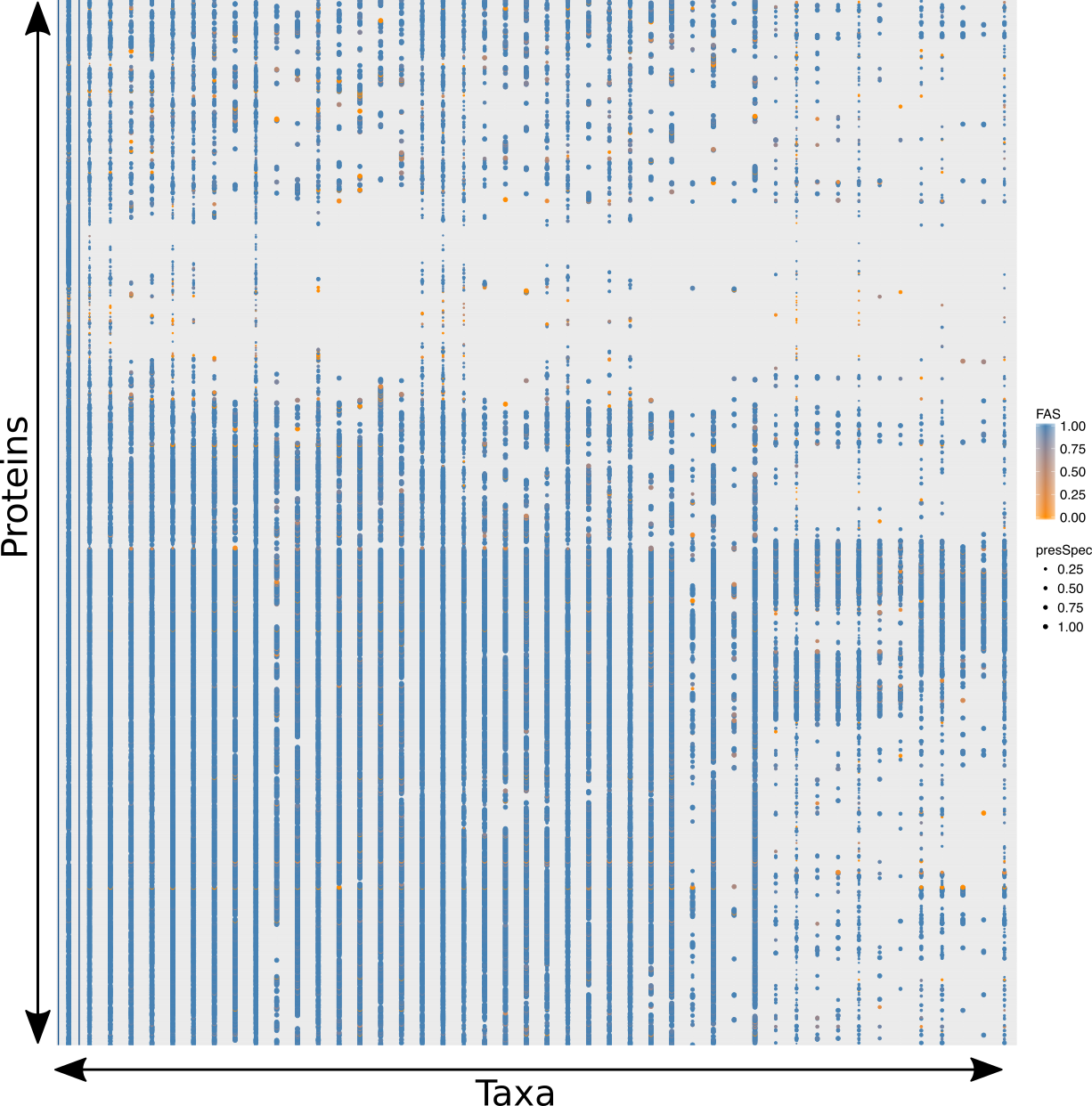


Figure A‑9: The full phylogenetic profile of 1605 microsporidia last common ancestor protein across 491 taxa grouped in phylum level. The color of the points denotes the FAS score between microsporidia and non-microsporidia protein. The size of the points depicts the percentage of species that have orthologs in each phylum.

Using PhyloProfile we estimated the evolutionary ages for microsporidia last common ancestor proteins. The estimation result can be seen in Figure A‑10. As expected, 94% of the proteins are as old as the last eukaryotic common ancestor, while only 3% are specific to microsporidia lineage.



Figure A‑10: Gene age estimation of 1605 microsporidia last common ancestor 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.

Out of 42 microsporidia specific proteins, only 6 have KO annotations (see Table A‑1) or 8 if combine HamFAS and BlastKOALA.

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

|  |  |  |
| --- | --- | --- |
| LCA protein | KO annotation | Description |
| OG\_1349 | K18592 | gamma-glutamyltranspeptidase / glutathione hydrolase / leukotriene-C4 hydrolase |
| OG\_1378 | K09485 | heat shock protein 110kDa |
| OG\_1710 | K04802 | proliferating cell nuclear antigen |
| 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 |

Table A‑2: KO annotation for 42 microsporidia specific proteins using BlastKOALA

|  |  |  |
| --- | --- | --- |
| LCA protein | KO annotation | Description |
| OG\_1087 | K17866 | diphthamide biosynthesis protein 2 |
| OG\_1378 | K09485 |  |
| 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 |  |
| OG\_2250 | K02896 |  |
| OG\_2280 | K02180 |  |

Gene Ontology terms (Ashburner et al. 2000) were assigned by Blast2GO v5.0.13 (Götz et al. 2008).

...

Discussion

Not out of our expectation, due to the compact genomes of extant microsporidia taxa, most of the proteins in the microsporidia last common ancestor should be evolutionary old. As 50% of the proteins are as old as the last universal common ancestor, another 44% proteins can be traced to the last common ancestor of all eukaryotes and 3% share the age with fungal clade, only 3% (or 42 proteins) are specific to microsporidia lineage. Those microsporidia specific proteins are still a mystery because of the poor functional annotation. There is no particular function or pathway that have been enriched by those proteins as has been seen from our KO and GO assignment analysis.

Conclusion

Functional annotation

Introduction

Proteins that are orthologous to each other are likely to have similar functions. The quality of orthology-based annotation transfer methods depends strongly on the accuracy of the ortholog prediction. Here we are introducing HamFAS, a robust annotation transfer pipeline based on feature-aware orthology inference. HamFAS has been shown to have higher sensitivity and comparable specificity in comparison to two state-of-the-art annotation tools KAAS and BlastKOALA from KEGG. A feature that makes HamFAS different than BlastKOALA and KAAS is the controllable ability of the annotation process. Users can choose different methods and threshold for increase or reduce the stringency of the annotation pipeline. Besides, HamFAS can be run locally through command lines. It provides a better solution for large-scale analysis than using online tools like KAAS and BlastKOALA.

Methods

HamFAS approach

We developed a novel approach named HamFAS to transfer KO annotations based on feature-aware orthology inference. Figure A‑11 demonstrates the pipeline of HamFAS.

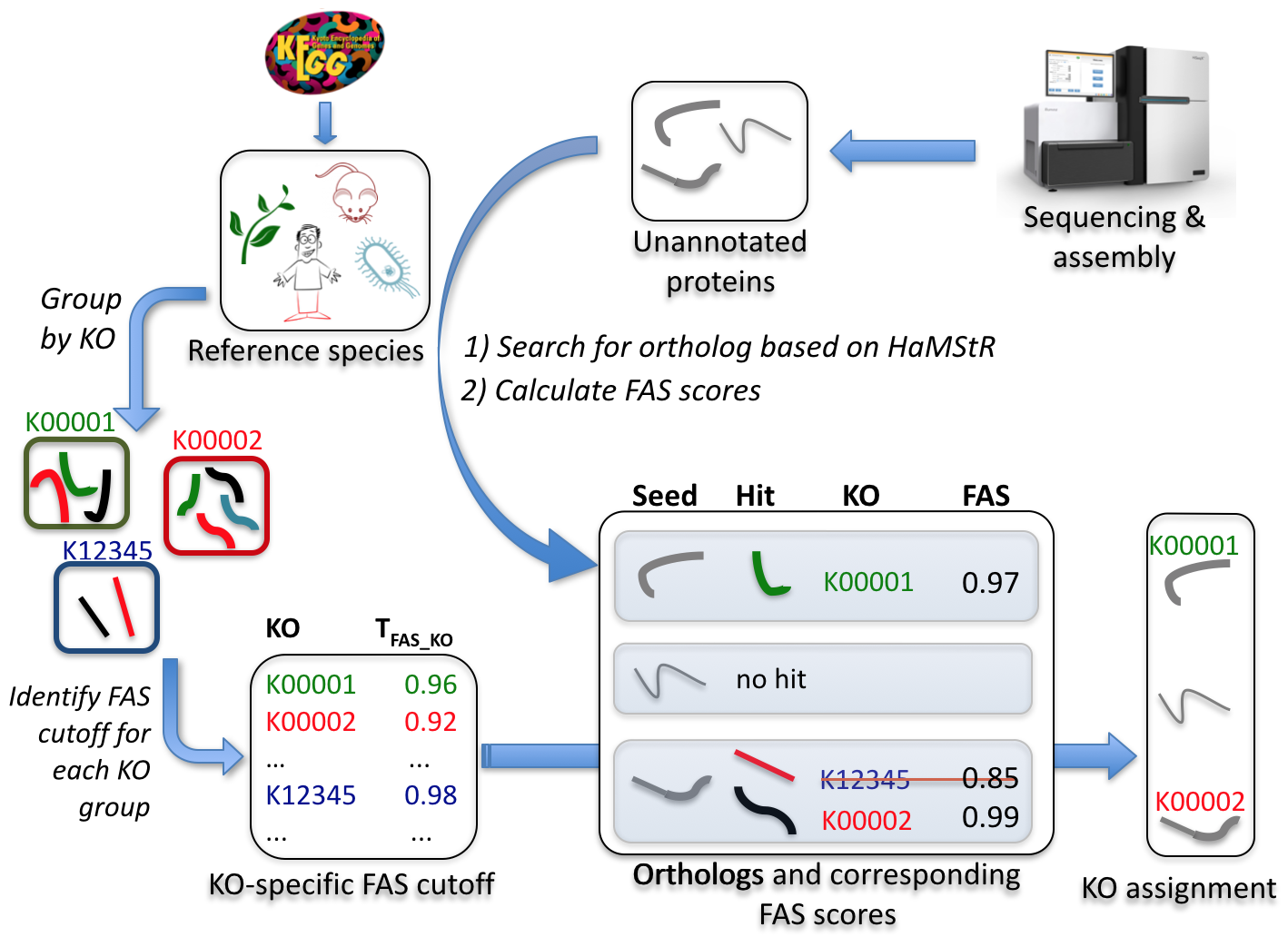


Figure A‑11: KO annotation transfer using HamFAS approach.

Protein sets of 30 manually KO-annotated reference species (Table A.5) have been downloaded from KEGG database. Pairwise FAS scores of all reference proteins within a KO group have been calculated. A group's mean FAS score serves then as a cutoff (TFAS\_KO) that must be exceeded to warrant transfer of its KO identifier to the seed proteins.

Given a list of uncharacterized proteins (seed), we search for their orthologs in the reference species using HaMStR (with *-checkCorothologsRef*, *-rbh* options and *-hit\_limit=5*). FAS scores between seed proteins and their orthologs will be identified. If the calculated FAS scores is not smaller then the corresponding TFAS\_KO, the available KEGG identifiers of those paired orthologs will be transferred to the seed proteins.

Benchmarking HamFAS

We used *S.cerevisiae* as a test species to benchmark our approach HamFAS. The protein set of yeast has been obtained from KEGG containing 3457 KO-annotated and 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 KAAS and BlastKOALA.

We removed *S.cerevisiae* out of the reference species list for avoiding redundant information while doing 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.

The ortholog search has been also performed with different parameters to find the best settings for HaMStR (*-rbh*, *-checkCoorthologRef*).

Results

Distribution of TFAS\_KO

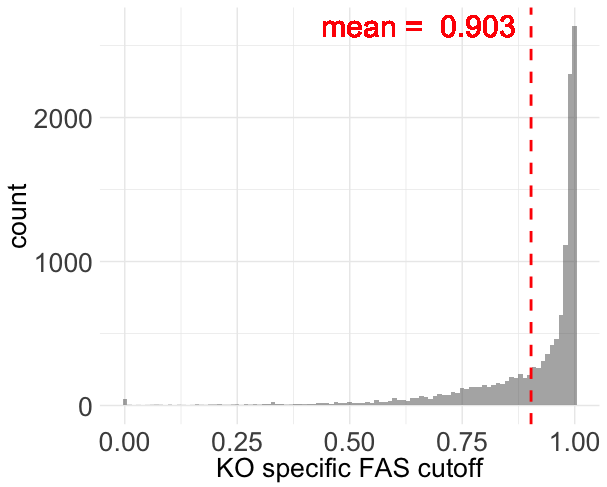


Figure A‑12: Distribution of TFAS\_KO for 12,748 KO groups

Figure A‑12 represents the distribution of all 12,748 TFAS\_KO values. 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‑13 shows 2 examples, the FAS scores distribution of K00542, which represents low TFAS\_KO group, and K07888, which represents high TFAS\_KO group.

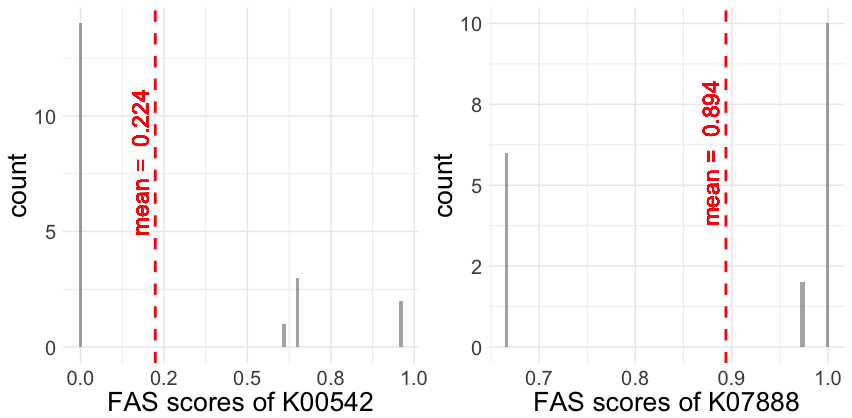


Figure A‑13: FAS score density of KO group K00542 (left) and K07888 (right)

In ortholog group K00542 (guanidinoacetate N-methyltransferase), only one protein member (rat rno:25257) has one 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 (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

3457 KO-annotated yeast proteins

With this data set, we tried to evaluate the accuracy of HamFAS in comparison to KAAS and BlastKOALA by calculating the recall, precision and F1 score.

recall = TP / (TP + FN)

precision = TP / (TP + FP)

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

Table A-3 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-3: 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 checking the ortholog prediction result obtained by HaMStR, we evaluated the annotation transfer again using only orthologs that are supported by both HaMStR and InParanoid. Predicted KOs from HamFAS of 188 yeast proteins has been removed after filtering based on InParanoid's orthologs. It leads to the decrease of recall and F1-score. However, the precision is not affected (see Table A-3). FAS scores of unsupported orthologs are slightly smaller than the ones of supported orthologs, with mean score of 0,918 and 0,988 respectively (see Figure A‑14).

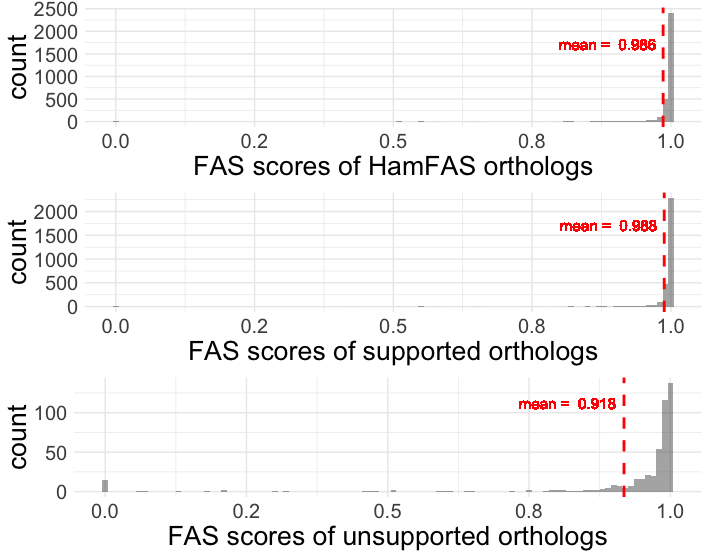


Figure A‑14: FAS score distribution of all HamFAS orthologs, only supported orthologs and unsupported orthologs

For a more detailed comparison between 3 approaches, we compare the fractions of proteins annotated by HamFAS, BlastKOALA and KAAS. 85,6% of the seed proteins has been annotated by all 3 approaches (Figure A‑15).

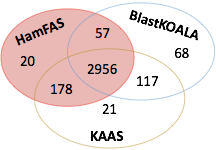


Figure A‑15: Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

There is a small difference between the KEGG identifiers annotated by each approach, which is shown in Table A-4 below.

Table A-4: Compare KEGG identifiers annotated by HamFAS, BlastKOALA and KAAS. Number 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 |

Although those KEGG identifiers are different, most of them are "synonymous" KOs. 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, etc.

Some examples of synonymous KOs:

* 1 KO is very general described (putative ABC transport system ATP-binding protein) while the other is specific (phospholipid/cholesterol/gamma-HCH transport system ATP-binding protein).
* Synonym/Alternative name: "septin" and "sporulation-regulated protein 3" (also septin); or "tristetraprolin" (ZFP36) and "butyrate response factor 1" (ZFP36L1).
* Involved in the same process: "cleavage stimulation factor subunit 2" and "polyadenylate-binding protein 2" are involved in 3-end formation of pre-mRNAs

3158 un-annotated yeast proteins

HamFAS could annotate 257 proteins, in which 164 proteins are HamFAS specific (HamFAS-only annotated proteins) (see Figure A‑16). In comparison to 150 and 116 annotated proteins from KAAS and BlastKOALA, HamFAS has annotated more proteins than BlastKOALA and KAAS (257 proteins versus 116 and 150 proteins, respectively)~~.~~

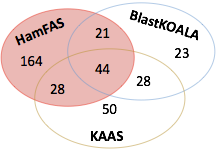


Figure A‑16: Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

Here the interesting part is the HamFAS-only proteins. So what are the differences of those proteins in comparison to others?

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 best specificity, HamFAS also shows the highest sensitivity in comparison to BlastKOALA and KAAS with the highest number of proteins that could be annotated.

Was HaMStR so inclusive to include many false positive orthologs?

We also compared the orthology search of HamFAS with InParanoid. After removing ortholog pairs that are not predicted by InParanoid, 150 out of 257 proteins still can be annotated, 55 of them belong to HamFAS-only annotated proteins. It proved the reliability of the orthology assignment of HamFAS-only proteins.

Are HamFAS-only proteins short and uninformative?

We compared the sequence length and the informative content of protein domains between HamFAS-only proteins and proteins annotated by at least 2 approaches including HamFAS and KAAS and/or BlastKOALA.

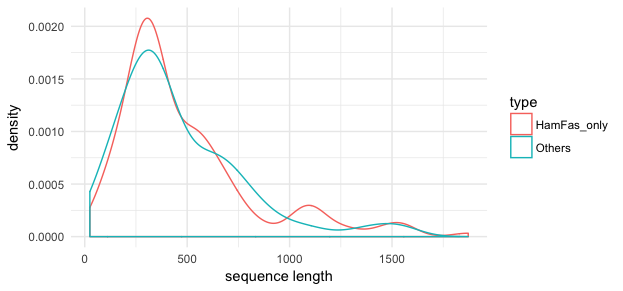


Figure A‑17: Length distribution of HamFAS-only proteins and others

Figure A‑17 and Figure A‑18 show no clear difference between those 2 protein sets. 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 (one Pfam domain that leads to the high FAS score).

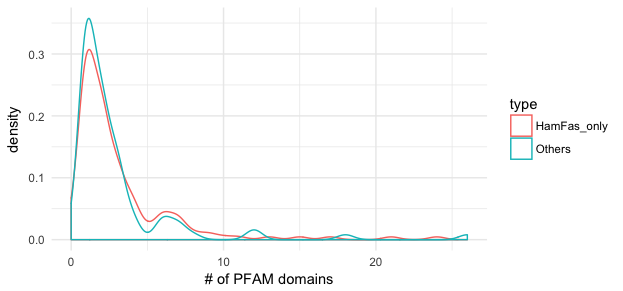


Figure A‑18: Number of Pfam domains distribution of HamFAS-only proteins and others

The distribution of FAS scores of all HamFAS orthologs in comparison to HamFAS-only orthologs shown in Figure A‑19 also confirms the rich domain annotations of HamFAS-only proteins.

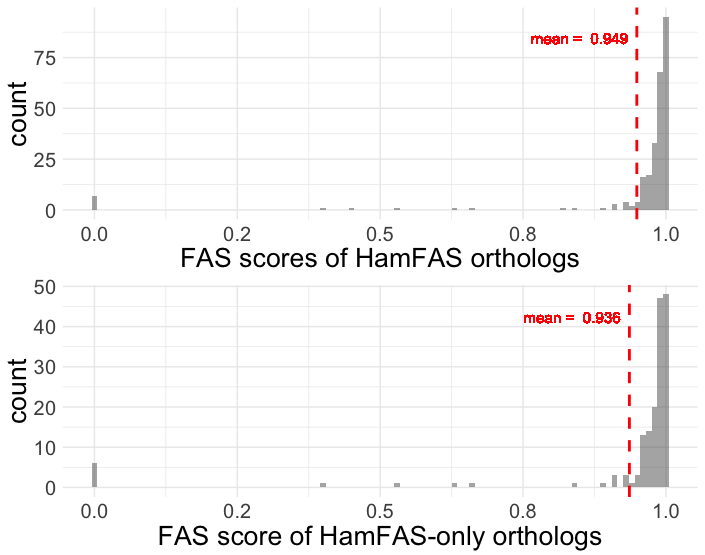


Figure A‑19: FAS score distribution of all HamFAS orthologs and HamFAS-only orthologs

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

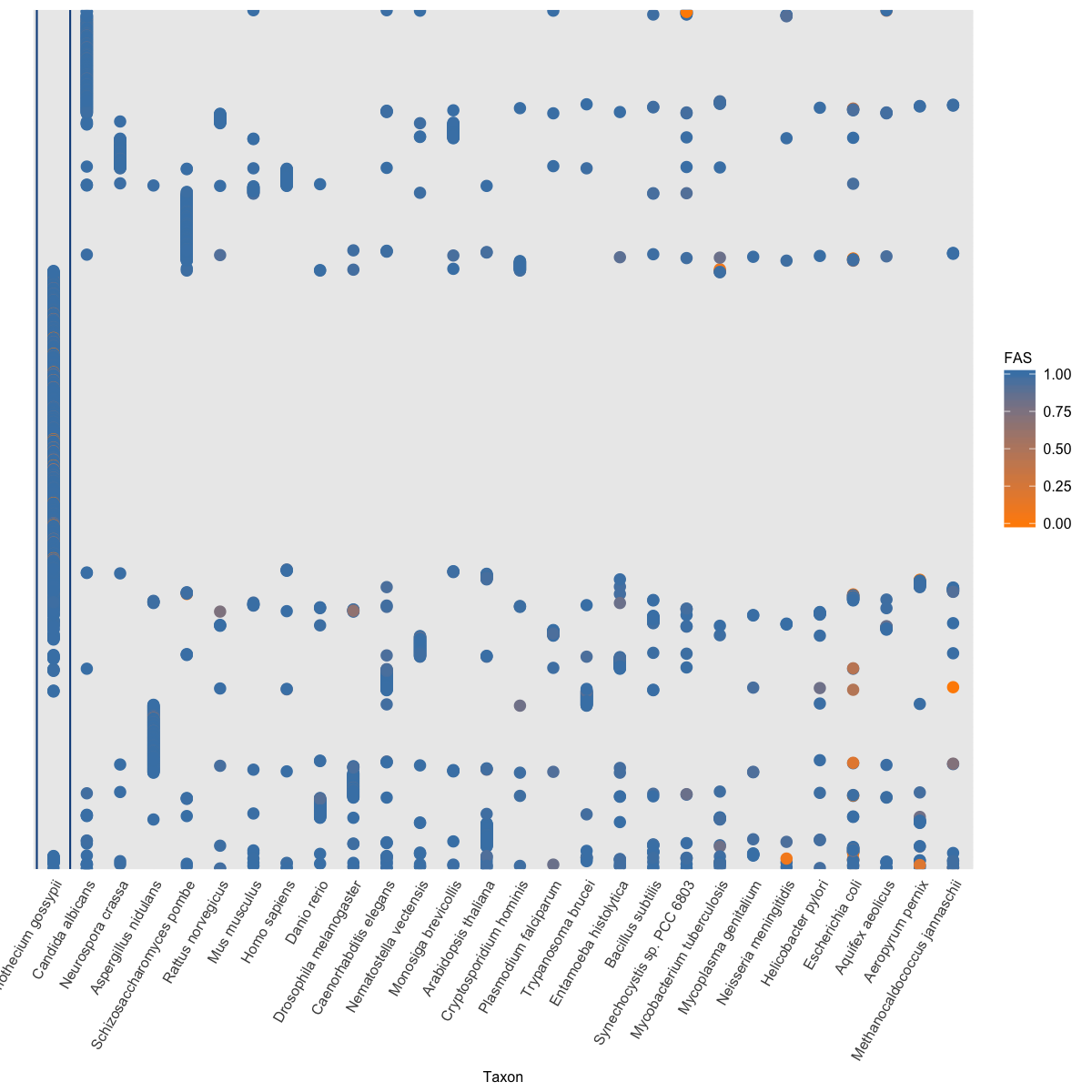


Figure A‑20: Phylogenetic profile of KO-annotated proteins

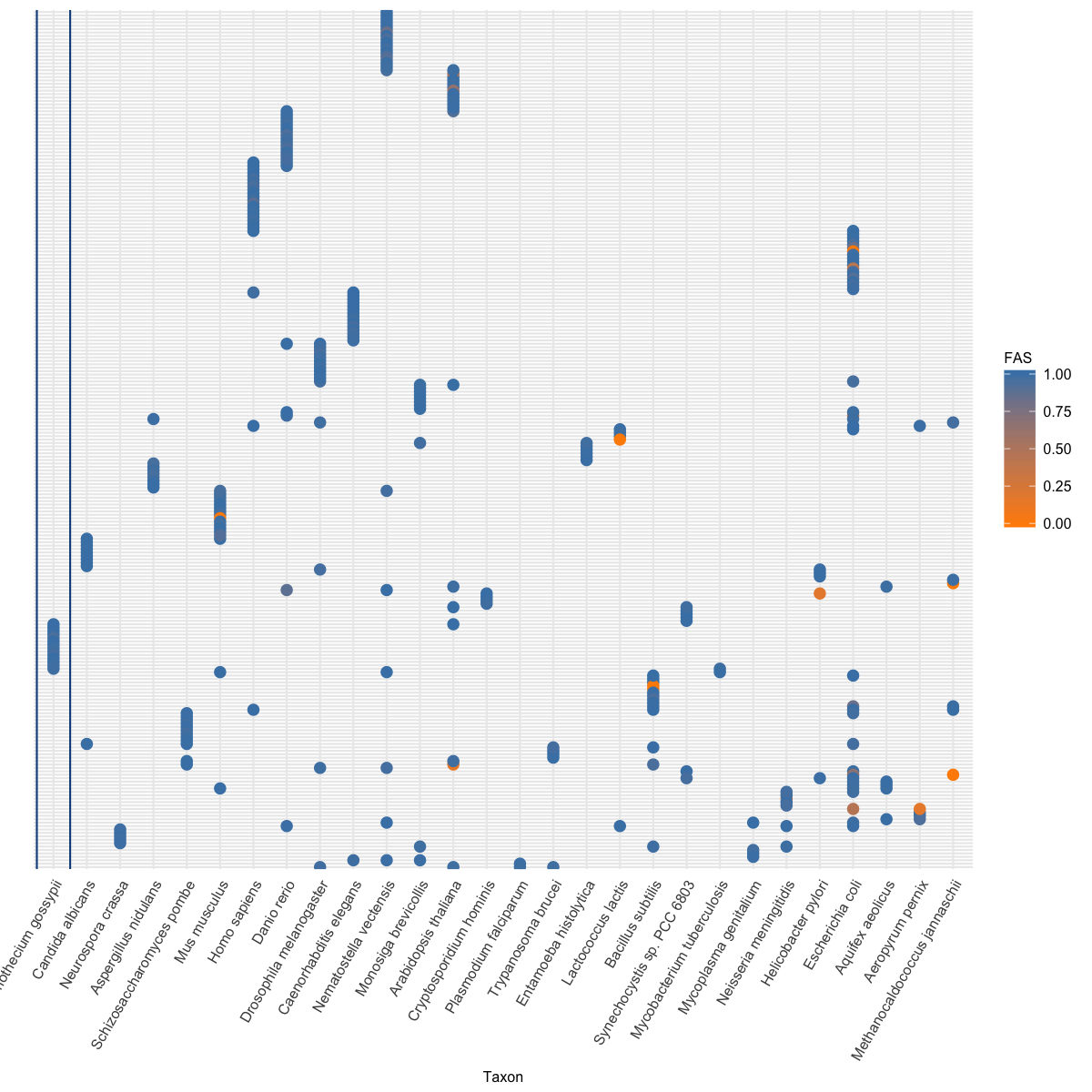


Figure A‑21: Phylogenetic profile of un-annotated proteins

Figure A‑20 and Figure A‑21 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‑22). 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.**).

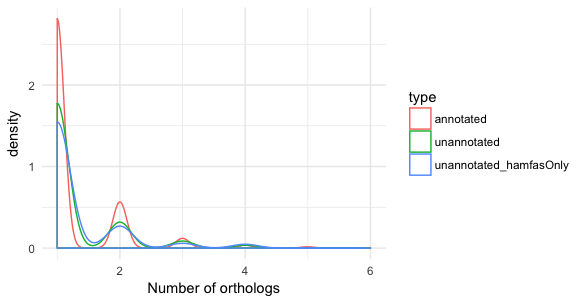


Figure A‑22: Distribution of number of orthologs for KO-annotated, un-annotated and HamFAS-only protein set

Do the annotations of HamFAS-only proteins come from distantly related species?

We checked for the origin of the annotations (i.e. the origin of reference orthologs) for all un-annotated proteins and compared with annotated set (Figure A‑23).

Figure A‑23: Origin of KO-annotations for annotated, un-annotated proteins and HamFAS-only proteins of un-annotated set

As expected, 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 large amount of annotations for un-annotated proteins originate from eukaryotes taxa (78%), there are still 22% (or 27% in case of HamFAS-only proteins) annotations are from distantly related taxa.

Analyzing the phylogenetic profile of proteins annotated by archaea and bacterial orthologs in Figure A‑24 and Figure A‑25, we can see that there is no difference between the HamFAS-only proteins and proteins that are annotated by both HamFAS and at least one other approach (HamFAS + BlastKOALA, HamFAS + KAAS or HamFAS + BlastKOALA + KAAS).

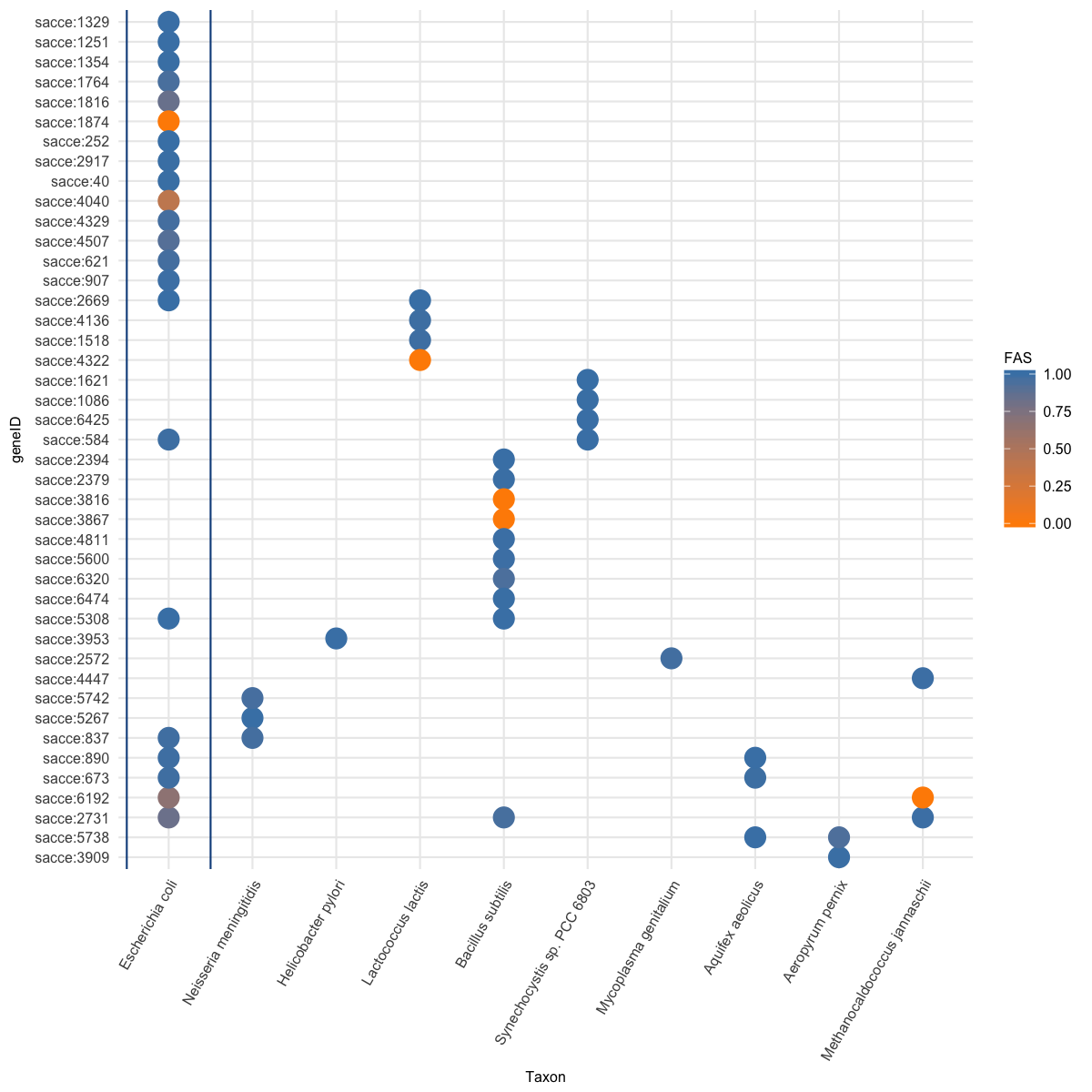


Figure A‑24: Phylogenetic profile of 44 HamFAS-only proteins that annotated based on archaea and bacterial orthologs.

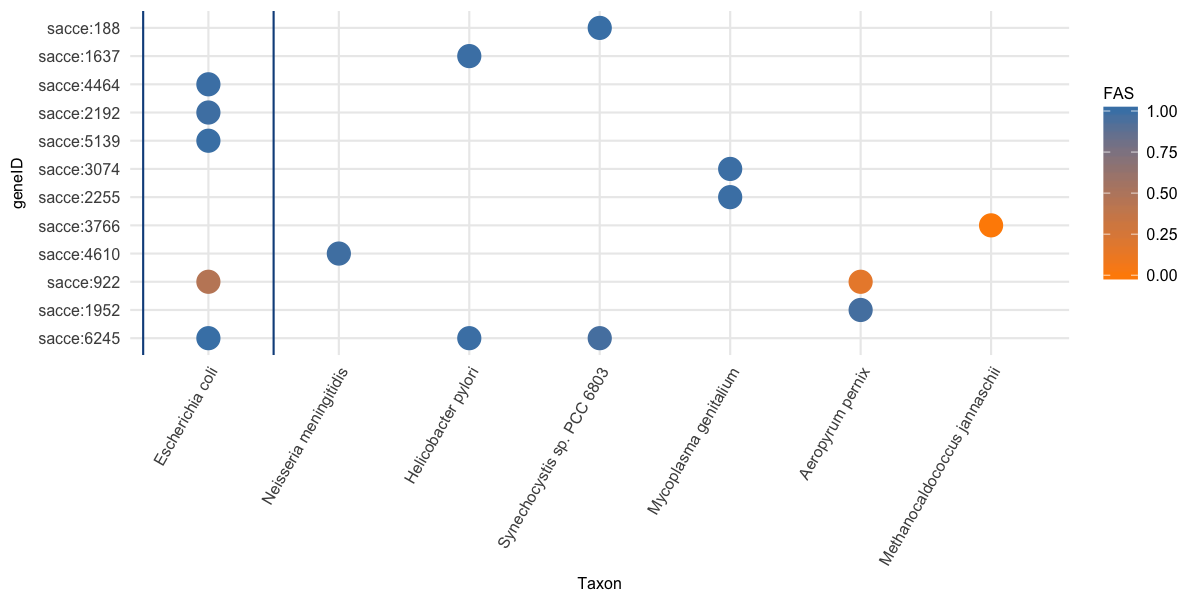


Figure A‑25: Phylogenetic profile of 12 un-annotated proteins that annotated by HamFAS and at least one other approach (BlastKOALA and/or KAAS), where their annotations originate from archaea or bacteria reference taxa.

How does the annotation result change by removing annotations from archaea and bacterial orthologs?

We filtered the annotations that originate from archaea and bacterial orthologs from both KO-annotated and un-annotated protein sets.

Table A-5: Recall, precision and F1-score of filtered HamFAS in comparison to HamFAS, BlastKOALA and KAAS by applying on KO-annotated yeast proteins.

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

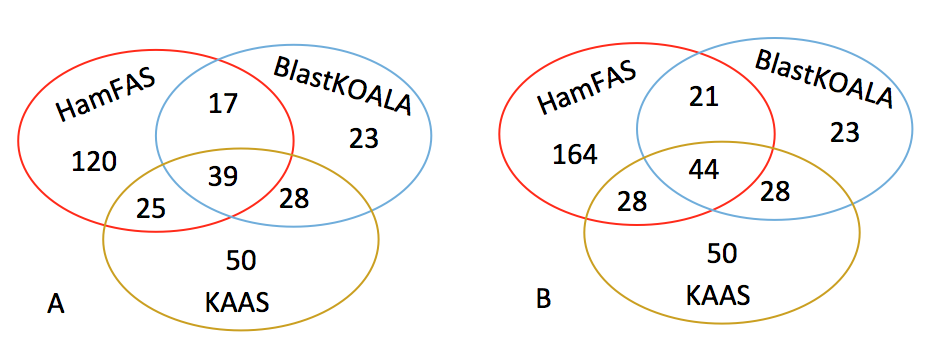
Table A-5 shows a slightly increase in precision and F1-score of filtered HamFAS in comparison to original HamFAS due to a small number of annotations obtained from distantly related taxa.

Figure A‑26: Fraction of proteins annotated by BlastKOALA, KAAS and filtered HamFAS (A) or original HamFAS (B)

In Figure A‑26 we observe a decrease of the number of proteins annotated by HamFAS. However there are still a large amount of proteins that are annotated only by HamFAS (120 proteins) in comparison to BlastKOALA (23 proteins) and KAAS (50 proteins).

Are annotated proteins involved in PPI networks or KEGG pathways?

We analyzed the connectivity of annotated proteins and the obtained KOs by calculating the node degree of those proteins in yeast protein-protein-interaction (PPI) networks and the occurrence of the annotated KOs in KEGG pathways. PPI data are retrieved from Yeast Interactome Project (http://interactome.dfci.harvard.edu/S\_cerevisiae/) and STRING database (https://string-db.org).

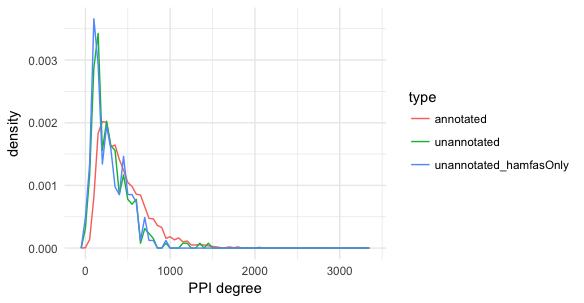


Figure A‑27: The PPI degree distribution of 3 protein sets

Figure A-27 shows the distribution of PPI degree of KO-annotated, un-annotated and HamFAS-only proteins inside un-annotated set. 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). 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.

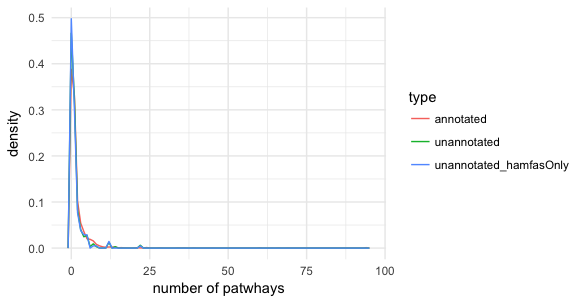


Figure A‑28: Distribution of the number of pathways in which annotated KOs are involved

For the annotated KOs, we calculate the number of pathways in which those KOs are involved. All 3 data sets show the same trend in Figure A-28, that 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%).

Are new annotations from HamFAS meaningless?

About 50% of KOs annotated only by HamFAS belongs to KEGG's pathways. Figure A-29 shows the distribution of those KOs in different pathway categories.

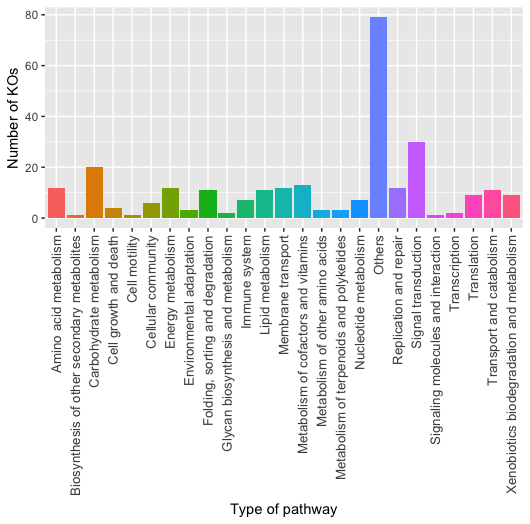


Figure A‑29: The numbers of HamFAS-only KOs distributed into different pathway categories

29 yeast pathways are further complemented by new KOs from HamFAS. (See appendix)

Why did BlastKOALA and KAAS fail to annotate HamFAS-only protein?

One reason could be, that the orthology prediction approaches used by KAAS and BlastKOALA are not as sensitive as HaMStR. The second reason is due the their "secret" filter criteria to select candidate for doing annotation transfer.

We have observed, that there are in total 86 predicted KOs that are common between annotated and unannotated data sets. In which 44 cases are also supported by InParanoid. Examples:

**K00077:** sce:4314, B.subtilis:BSU15110, B.subtilis:BSU14440 (KEGG's representative sequences)

HamFAS: sce:6474 (unannotated protein) is orthologous with B.subtilis:BSU14440

Inparanoid: 2 separate OGs: (sce:4314, B.subtilis:BSU15110) and (sce:6474, B.subtilis:BSU14440)

**K00799:** ath:AT1G02930,..., ath:AT2G30870,..., sce:5364, sce:1884

HamFAS: sce:2310 - ath:AT1G02930,...

(sce:1884 has no ortholog with *A.thaliana* according to InParanoid)

**K00877:** S.pombe:4570, S.pombe:875, S.pombe:1336, sce:1877, sce:997

HamFAS: sce:487 - S.pombe:1336

Those proteins have been probably either not predicted as orthologs or discarded after filtering through KEGG annotation pipeline.

Conclusion

The ability of identifying distantly related orthologs of HaMStR leads to the result that more proteins have been annotated by HamFAS than BlastKOALA or KAAS. There is no strong evidence to distinct the difference between HamFAS-only proteins and proteins that are annotated by both HamFAS and other approaches. 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. The principle factor that affects the annotation result is the accuracy of orthology assignment method. This HamFAS approach could be supported more by the analysis of QfO from Holger.

Metabolic pathway analysis of microsporidia last common ancestor proteins

Introduction

Metabolic analysis of microsporidia is still a challenge due to their obligate intracellular growth and short lifetime of its purified spores (Keeling & Fast 2002).

Methods

We used HamFAS approach to do KO annotation for 1605 microsporidia last common ancestor proteins. HaMStR was used for ortholog search between microsporidia last common ancestor (seed) and KEGG reference species (reference). Since one microsporidia last common ancestor protein is represented by an orthologous group of several microsporidia proteins, we assigned the representative FAS score for each reference protein as the max score that protein can archive when compare with all microsporidia proteins in the corresponding orthologous group. This representative max FAS score will be then compared with the TFAS\_KO in order to decide if the annotated KO of the reference protein can be transferred to the microsporidia last common ancestor protein.

Beside the complementary FAS scores to the orthology assignment, we also calculated the patristic distance between the reference protein and microsporidia protein to use it as a confident value for the annotation transfer. The distance between a reference protein to the microsporidia last common ancestor protein is the minimum distance between that reference protein to all microsporidia proteins in the corresponding orthologous group. The distances in one orthologous group are normalized to a range of [0,1] by the formula (currentDist - minDist)/(maxDist - minDist).

The KO-annotated microsporidia last common ancestor proteins were then mapped to KEGG pathways. The result was compared with 4 extant microsporidia species E.cuniculi, E.hellem, E.intestinalis, N.ceranae, which are available in KEGG database, and with S.cerevisiae, as an example for free-living organism. The annotations for E.cuniculi, E.hellem, E.intestinalis, N.ceranae and S.cerevisiae were obtained directly from KEGG.

Results

How many proteins are annotated?

Compare network diameters (or all graph properties as shown in table 2) between parasite and non-parasite species (diameters of parasites' networks are NOT smaller than those of non-parasites (Nerima et al. 2010) ), may be only for core metabolism (907 or 876 nodes - NO, different in type of nodes. We have enzymes as nodes, they use educt-product as nodes).

Pathway enrichment result of microsporidia last common ancestor is shown in Figure A‑30.



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

Figure A‑31 shows the fractions of proteins take part in different pathway categories of microsporidia last common ancestor in comparison to other extant species. microsporidia last common ancestor 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.

A more detail of the mapped pathways and number of proteins for each pathway is shown in Figure A‑32. In general, microsporidia last common ancestor 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. It suggests that the microsporidia last common ancestor already have a parasitic-living style.



Figure A‑32: Number of proteins of each taxon (green for microsporidia last common ancestor, 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.

*get some pathways where there are big difference between LCA and extant species -> show connectivity network -> phylogenetic profiles*

*not work: for example, studies said that micros don't have glycoglysis/glucogenesis but all extant micros in KEGG have this pathway.*

*=> LCA annotation is mostly from micros, how confident we are with the annotation? how can we identify genes that lost in LCA?*

*find papers about metabolic pathways of micros and check if LCA missing the mentioned proteins?*

In many eukaryotes, the major facilitator superfamily (MFS) and ATP binding cassette (ABC) transporters (Table S14) are present as multiple subfamilies with a broad range of substrate specificities [61,62]. By contrast, T. hominis has retained only putative sugar-, folate- and peptide- transporting MFS transporters (Table S14, Figure S12), and only two groups of ABC transporters: including those which are typically found in mitochondria (ABCB and ABCG [63,64]; Figure S13). Comparison of the predicted transporters for T. hominis with other microsporidians (Figure S11, Table S14) suggests that microsporidians retain a similar repertoire of core transport proteins. (Heinz et al. 2012)

MAIN PATHWAYS: glcoglysis/glucogenesis, Krebs cycle (TCA), pentose phosphate, purine/pirimidine, amino acid metabolism (Nerima et al. 2010).

In all such species, many small molecule biosynthetic pathways are completely absent, and core carbon metabolism is limited to glycolysis, the pentose phosphate pathway, and synthesis and degradation of the storage carbohydrate trehalose—even the smallest E. intestinalis genome has the full complement of genes for these pathways. (Keeling & Corradi 2011)???

Core carbon metabolism in T. hominis (Figure S1) appears to mainly comprise the interconnected pathways for glycolysis, the pentose phosphate pathway, and trehalose biosynthesis and catabolism. With the exception of E. bieneusi, which appears to have lost all three pathways [11,55], these pathways are also strongly conserved in the other microsporidians (Heinz et al. 2012)

Like many parasites [58], T. hominis has lost the ATP-expensive pathways for the de novo biosynthesis of inosine 59-phosphate (IMP) and for uridine 59-phosphate (UMP); the starting points for the biosynthesis of purines and pyrimidines for DNA and RNA biosynthesis (Heinz et al. 2012)

Altogether, these studies confirmed the suspected lack of TCA cycle, showed a requirement for ATP in the sustaining media (buttressing suspicions that these parasites probably import ATP from their hosts), determined that the parasites produce lactic acid and pyruvic acid, and demonstrated the presence of several enzymes involved in glycolysis, the pentose-phosphate pathway, as well as trehalose synthesis and degradation (22, 84, 102). These studies indicate that microsporidia have retained the glycolytic pathway and suggest that they probably use extended glycolysis. (Keeling & Fast 2002)

Lack TCA, oxidative phosphorylation, electron transport, glycosis (Keeling 2009); have several atp transport (Vivarès & Méténier 2001; Keeling 2009); have hsp70 and pyruvate dehydrogenase (Fast & Keeling 2001; Keeling 2009) => LCA had mitochondria

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

lack of mitochondria and peroxisomes (Agnew et al. 2003; Fast & Keeling 2001)

According to (Fast & Keeling 2001; Agnew et al. 2003; Keeling & 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. The extant microsporidia species lack electron transfer chains, oxidative phosphorylation, and the tricarboxylic acid (TCA) cycle, just like other amitochonriate species. They "break down glucose using the glycolytic pathway, which is like that of other eukaryotes except that phosphofructokinase is pyrophosphate-dependent rather than ATP-dependent" (NEED TO BE REPHASED). They will replace pyruvate dehydrogenase complex (PDH) by pyruvate ferredoxin oxidoreductase (PFOR) in order to convert pyruvate into acetyl-CoA and produce NADH. We could not find any KOs of 4 PFOR subunits (α, β, γ, δ) in microsporidia LCA, however 3 out of 4 components of PDH were found instead, i.e. pdhA (OG\_2283, EC=1.2.4.1) and pdhB (OG\_2084, EC=1.2.4.1) of E1 component, and E3 (DLD) component (OG\_3281, EC=1.8.1.4) (E2 DLAT, EC=2.3.1.12 not found). (CHECK AGAIN WITH NEW DATA). Note that E1 is also be found in N.locustae (Fast & Keeling 2001) and Encephalitozoon (Katinka et al. 2001). Figure A‑33 shows the mapped microsporidia last common ancestor proteins into the reaction converting pyruvate into acetyl-CoA.

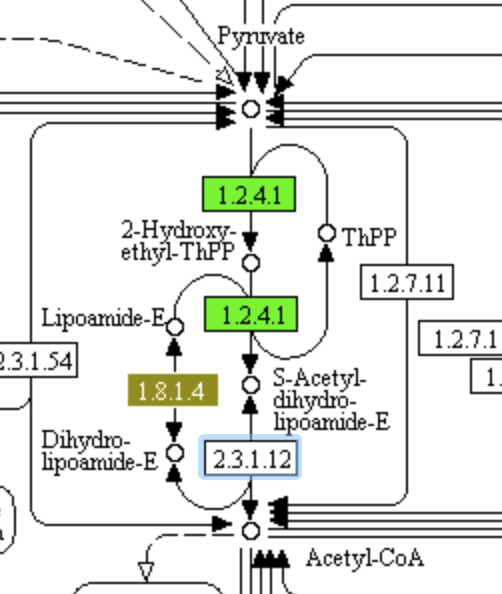


Figure A‑33: (NEED TO BE REDRAWN) The reaction Pyruvate -> Acetyl-CoA with help of pyruvate dehydrogenase complex (PDC). Proteins present in microsporidia last common ancestor are highlighted.

The role of E1 component and the reason for the presence of hsp70 genes in extant microsporidia is still unclear (Fast & Keeling 2001). But the presence of two subunits of E1 and the component E3 in the last common ancestor emphasizes the origin of mitochondria in the microsporidia ancestor.

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).

phosphofructikinase is pyrophasphate-dependent rather than atp-dependent; pyruvate metabolism is perhaps different; core metabolism (Keeling & Fast 2002)

spores of micros contain large amount of trehalose (primary carbohydrate-storage in lichens) => used for ATP production via glycolysis. host species convert sugar into carbohydrates (glucose and trehalose) & lipid => exploited by microsporidia (because they lack oxidative respiration by mitochondria) (Agnew et al. 2003).

Glycolysis and trehalose catabolism are thought to be important for microsporidian spore survival and germination [2,56], and, consistent with this hypothesis, neutral trehalase and most of the proteins needed for glycolysis, the alternative oxidase and glycerol- 3-phosphate dehydrogenase, were represented in the proteomics data (Heinz et al. 2012)

lca uptakes atp from host or can produce by itself? (find reactions that produce ATP and search for their enzymes in lca) Because (Méténier & Vivarès 2001) said that micros mostly uptake atp from host.

enzymes for trehalase activity (Méténier & Vivarès 2001) which are found only in Nosema apis (Vandermeer & Gochnauer 1971) => also found in LCA? (carbohydrate and energy metabolism)

Micro- sporidian genomes lack the enzymes needed for the de novo synthesis of nucleotides [4,6] (Figs 2 and 3) including phosphoribosyl pyrophosphate (PRPP) synthase [4], which makes the sub- strate PRPP that is required for the activation of ribose-5-phosphate for both purine and pyrimidine de novo pathways [16,17]. The loss of this biosynthetic function alone means de novo synthesis is not possible. Other key enzymes are also absent [18] (Fig 3), including IMP cyclohydrolase that makes inosine monophosphate (IMP)—the first purine nucleotide in the de novo pathway—and UMP synthase [18], which makes the first pyrimidine nucleotide, uri- dine monophosphate (UMP). Fig 2. Nucleotide biosynthesis in microsporidia. (Dean et al. 2016)

Phylogenetic analyses [18,37,38] suggest that a single NTT gene was probably acquired by horizontal transfer from bacteria into the microsporidian common ancestor. This was followed by lineage-specific gene duplications to generate the multiple copies of NTT genes found in contemporary microsporidian genomes [4,6,18]. For example, the microsporidia T. hominis, E. cuniculi, and E. bieneusi have four NTTs while Spraguea lophii has six [39] (Dean et al. 2016) => find this!! (and check if the annotation come from bacteria)

can replace extant microsporidia by contemporary microsporidia

Discussion

Conclusion

# Appendix

Tables

Table A‑1: Eleven extant microsporidia species

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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: 24 taxa used for extent the initial homologous groups

|  |  |  |  |
| --- | --- | --- | --- |
| 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 microsporidia last common ancestor 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* |

