# The estimation of the microsporidian last common ancestor protein set

Proportion of orthologous and lineage specific proteins (or the reduction and expansion in microsporidian genome)

We analyzed the proportion of orthologous and lineage specific proteins in 11 microsporidia species.

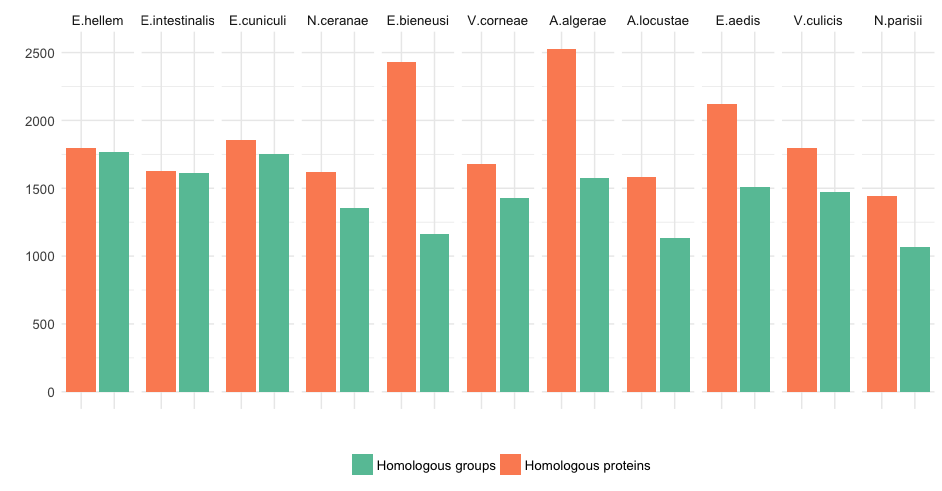


Figure 1‑1: The distribution of number of homologous proteins (orange) and number of homologous groups (green) in each microsporidia taxon.

Figure 1‑1 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 1‑2 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 and Koonin, 2004).

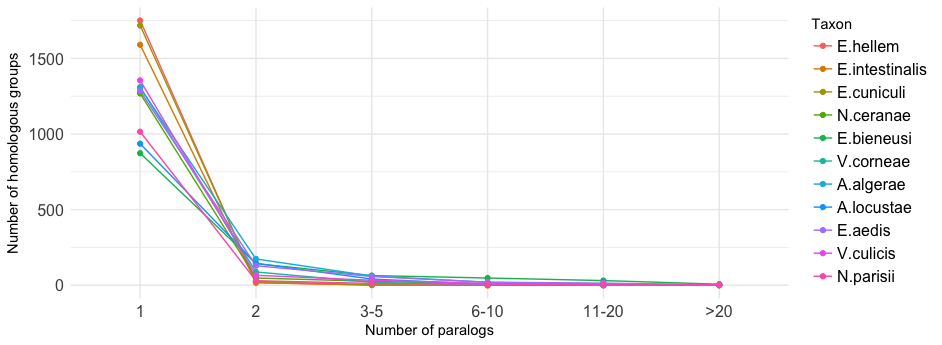


Figure 1‑2: The distribution of number of homologous groups as a function of number of in-paralogs. Colors denote different microsporidia taxa.

# HamFAS:

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

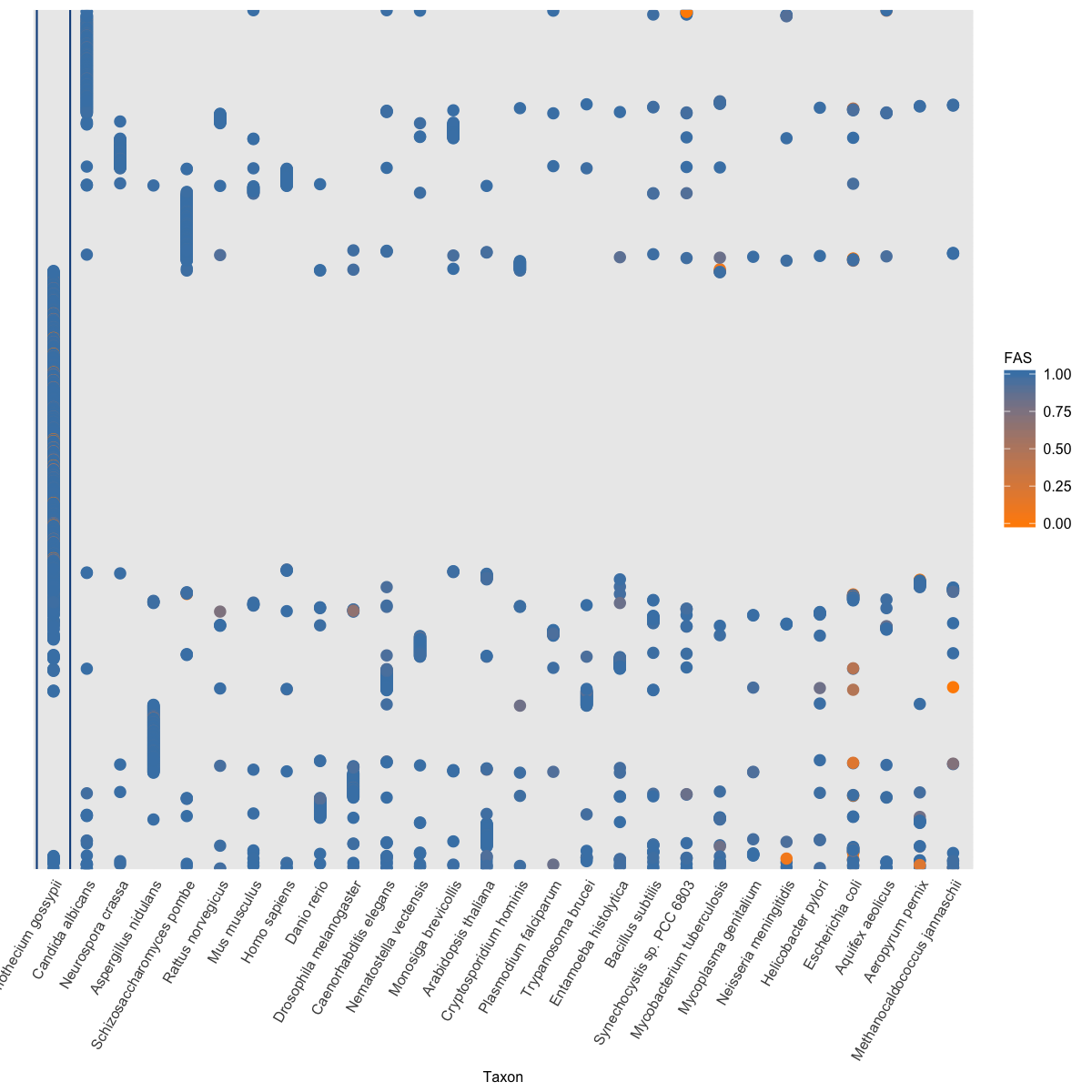


Figure 2‑1: Phylogenetic profile of KO-annotated proteins

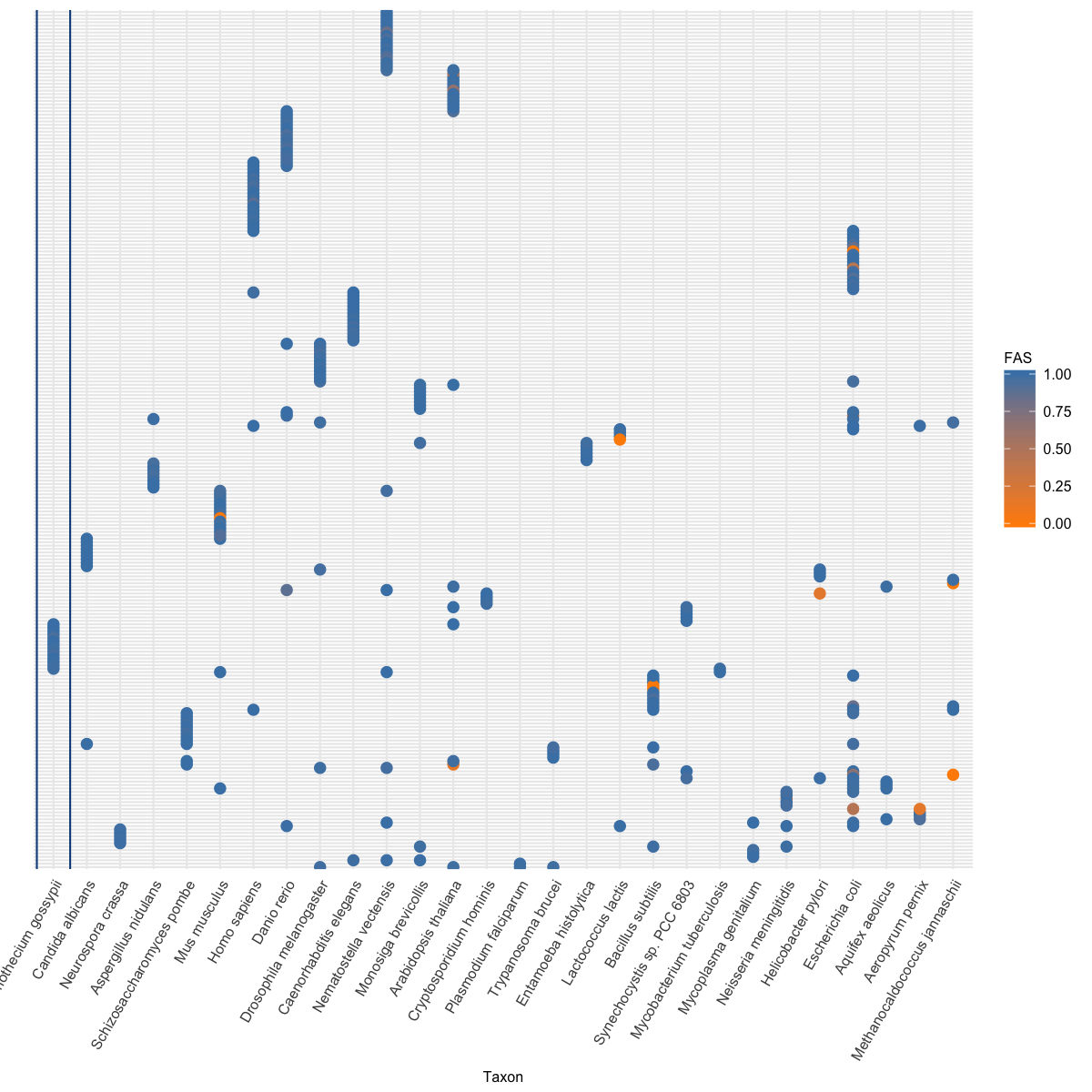


Figure 2‑2: Phylogenetic profile of un-annotated proteins

Figure 2‑1 and Figure 2‑2 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 2‑3). 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.**).

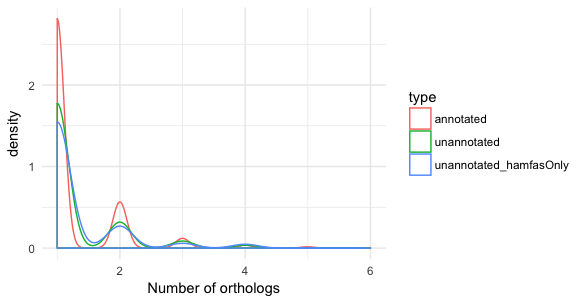


Figure 2‑3: Distribution of number of orthologs for KO-annotated, un-annotated and HamFAS-only protein set

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.

# Metabolic pathway analysis

(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

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 3‑1: 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 3‑1 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 3‑2. 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 3‑2: 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 3‑1.

Table 3‑1: 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 |