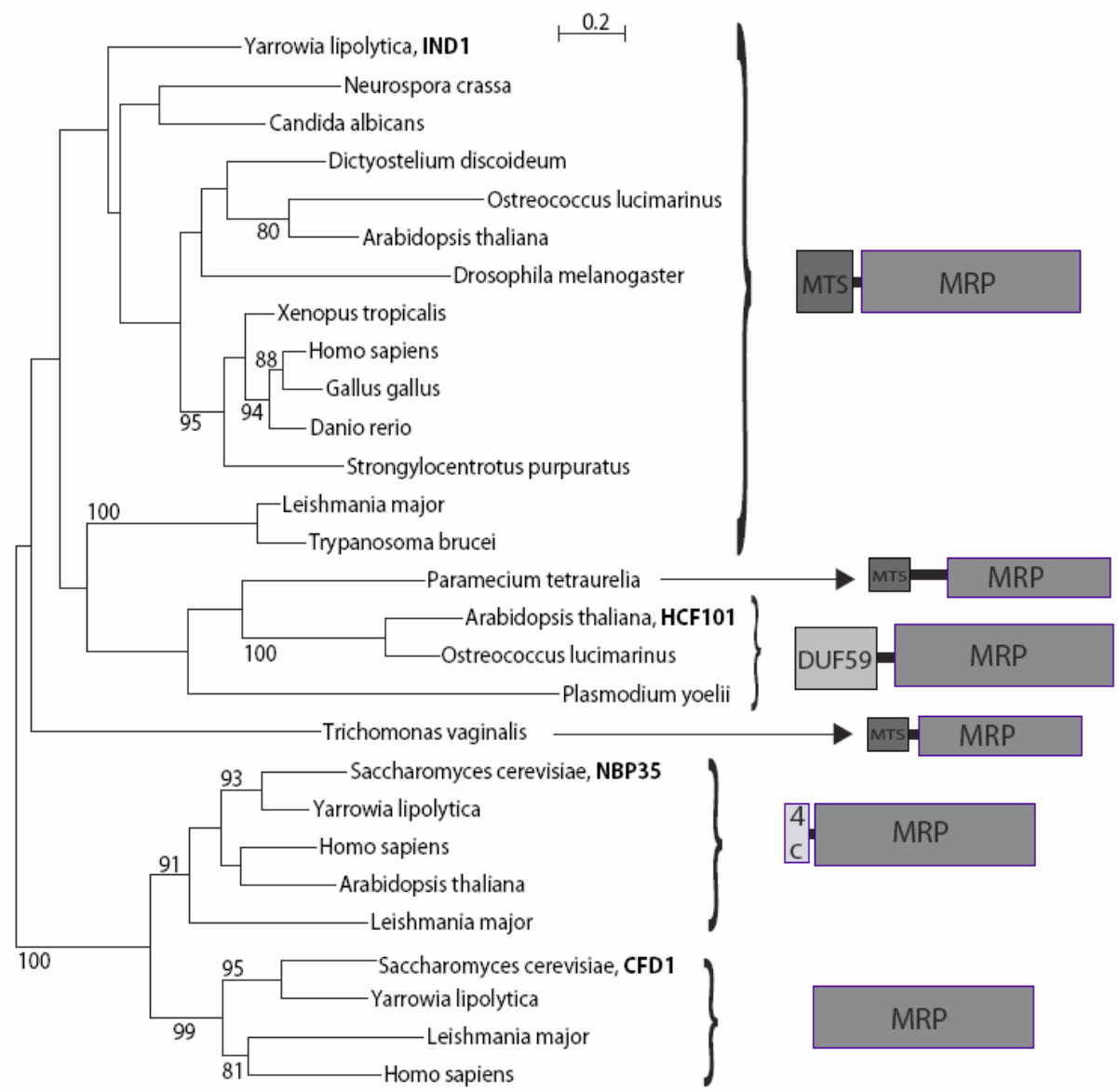
Supplementary Figure S1 - Phylogenetic relationships and domain compositions of *IND1*, *HCF101*, *CFD1* and *NBP35*.



Eukaryotic representatives of the *IND1*, *HCF101*, *CFD1* and *NBP35* subfamilies were assembled using PSI-Blast searches in completely sequenced genomes. The separation into the four different subfamilies is generally supported by the domain structure of the proteins. The metazoa, fungi, plants, algae, and *Dictyostelium* with complex I clearly have an orthologue of *IND1*, while the ones without do not have an orthologue of *IND1* (the fungi *Saccharomyces cerevisiae*, *Vanderwaltozyma polyspora*, *Kluyveromyces lactis*, *Ashbya gossypii*, *Candida glabrata* and *Schizosaccharomyces pombe*, the Amoebozoan *Entamoeba histolytica*, the Apicomplexa and *Giardia*). The phylogenetic positions of the *IND1*-like sequences from Euglenozoa (containing complex I) and from *T. vaginalis* (containing only two complex I proteins) is not well resolved. However, in terms of domain structure they are identical to *IND1* from *Y. lipolytica* and were considered orthologous.
The situation in the complex I containing ciliate *P. tetraurelia* is less clear-cut, as its *IND1* homolog appears more closely related to *HCF101* than to *IND1*. Nevertheless, similar to

IND1 and unlike the plastid-localized *HCF101*, the ciliate protein has a mitochondrial targeting signal, and the DUF59 domain has eroded (E-value > 0.1). Domain compositions were examined using SMART (Letunic et al., 2006) and the presence of a mitochondrial targeting signal (MTS) was examined using Mitoprot (Claros and Vincens, 1996), using a cut-off score of 0.8. The alignment was derived using ClustalX (Thompson et al., 1997). In the domain compositions the presence of the four cysteines (4C) at the N-terminus of *NBP35* is indicated. The phylogeny is based on the well-aligned region of the Mrp domain (Supplementary Figure S4). The tree was derived with PhyML (Guidon et al., 2005), using the default options: four gamma distributed rate categories and the WAG substitution model. Bootstrap values larger than 75/100 are indicated. The gene identifiers of the sequences used for the alignment and the tree are, from top to bottom: *Y. lipolytica*: 50547189, *N. crassa*: 68472597, *C. albicans*: 85097286, *D. discoideum*: 66803064, *O. lucimarinus*: 1145348579, *A. thaliana*: 15235067, *D. melanogaster*: 116008233, *X. tropicalis*: 62857965, *H. sapiens*: 157384956, *G. gallus*: 50748402, *D. rerio*: 157423523, *S. purpuratus*: 115774549, *L. major*: 157871966, *T. brucei*: 74025340, *P. tetraurelia*: 145487614, *A. thaliana*: 15230111, *O. lucimarinus*: 145355520, *P. yoelii*: 68525518, *T. vaginalis*: 123414978, *S. cerevisiae* 6321347, *Y. lipolytica*: 2912096, *H. sapiens*: 118572611, *A. thaliana*: 9758243, , *L. major*: 157868894, *S. cerevisiae*: 6322188, *Y. lipolytica: 2911744*, *L. major*: 157871001, *H. sapiens*: 6912540.

Claros, M.G. and Vincens, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem*. **241**, 779-786.

Guindon, S., Lethiec, F., Duroux, P. and Gascuel, O. (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res*. **33** (Web Server issue), W557-559.

Letunic I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J. and Bork, P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*. **34** (Database issue), D257-260.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **24**, 4876-4882.