Jason Jiang

Title: Characterizing functional consequences of reductive protein evolution in *Microsporidia*

Supervisor: Dr. Aaron Reinke

Authors: Jason Jiang, Rui (Jerry) Qu, Aaron Reinke

Summary

Microsporidia are an elusive and successful group of eukaryotic parasites, causing fatal illness in humans and harming economically important livestock. They are found in every major continent and parasitize a wide range of metazoans and protists. Microsporidia have undergone remarkable genomic adaptations to their host-dependent lifestyles, losing key metabolic genes and shortening their proteins after they diverged from fungi. In this study, I investigate the evolution of protein domain architectures of *Microsporidia* orthologs to *Saccharomyces cerevisiae* (yeast) and test whether these orthologs can replace function in knockout yeast strains. *Microsporidia* were revealed to undergo clade-specific patterns of domain loss, reflecting unique adaptations to different environments. Moreover, the *Nematocida parisii* ortholog to essential yeast gene *MPE1* was unable to rescue a temperature-sensitive strain for *MPE1*. This study provides further insight into the extent to which these parasites are host-dependent and reveals novel protein-level adaptations to their parasitic lifestyles.

Introduction

Microsporidia are a poorly understood group of obligate intracellular parasites, causing devastating illness in humans and economically important livestock (Wadi and Reinke, 2020). They are notable for their simplicity, being unicellular eukaryotes with reduced organelles (Wadi and Reinke, 2020). Despite their simplicity, they are successful parasites in every major continent, parasitizing a wide range of metazoans and protists (Murareanu et al., 2021). Microsporidia are believed to have evolved as a sister group to all fungi (Nakjang et al., 2013), being the earliest diverging clade (James et al., 2013). Since their divergence, they have undergone remarkable genomic adaptations to their parasitic lifestyles (Keeling and Fast, 2002). For instance, all microsporidia have large expansions of transporter gene families, which are used to steal host metabolites (Nakjang et al., 2013). Moreover, microsporidian genomes are highly compact and gene-dense (Corradi and Slamovits, 2011), with *Encephalitozoon intestinalis* having the smallest eukaryotic genome of 2.3 Mbp (Corradi et al., 2010). Due to their host-dependency, microsporidia have lost key metabolic gene families, with *Encephalitozoon cuniculi* lacking genes for core carbon metabolism (Keeling et al., 2010). Microsporidian proteins are also reduced, with microsporidian proteins being on average ~15% shorter than their yeast orthologs

(Katinka et al., 2001). However, it is unclear how functionally conserved these reduced microsporidian proteins are to their fungal orthologs.

Proteins consist of independently folding "domains" connected by amino acid linkers. Domains are the functional units of proteins, with linkers facilitating molecule binding and protein-protein interactions (Wang et al., 2011). The domain architecture of a protein is its linear arrangement of domains, from N to C-terminus (Forslund et al., 2011). As domains are the functional units of proteins, protein function and evolution may thus be understood through their domain architectures.

Moreover, the model organism *Saccharomyces cerevisiae* (yeast) has been used extensively for gene functional replacement studies (Kachroo et al., 2015; Kachroo et al., 2017; Laurent et al., 2020), where orthologs from other species are cloned and transformed into yeast to determine if they can functionally replace their cognate yeast genes. These studies have been facilitated by extensive research done on the yeast genome (Botstein et al., 1997) and readily available and experimentally tractable knockout yeast strains. For example, heterozygous knockout (hetKO) yeast strains are heterozygous for a G418 antibiotic resistance cassette replacing an essential gene (Giaever and Nislow, 2014) and temperature-sensitive (ts) strains have alleles for an essential gene that become non-functional at some restrictive temperature (Ben-Aroya et al., 2008; Li et al., 2011).

In this study, I take a computational and experimental approach to investigate the functional conservation of *Microsporidian* proteins to their fungal orthologs, using yeast as a model fungus. Computationally, I developed a bioinformatics pipeline to detect single-copy orthologs between *Microsporidia* species and yeast and compare orthologous domain architectures. This pipeline revealed a set of truncated and mostly yeast-essential orthologs between *Microsporidia* and yeast. Moreover, domain loss was highly prevalent, occurring in ~25% of *Microsporidia* orthologs to yeast. When possible, non-essential C-terminal domains were preferentially lost in *Microsporidian* orthologs, reflecting a functionally conservative strategy for protein reduction. Moreover, *Microsporidia* orthologs undergoing domain loss were enriched for broad metabolic functions, reflecting the reduced metabolic capacity and extensive host-dependency of *Microsporidia*. Furthermore, domain losses occurred in highly clade-specific manners in *Microsporidia*, potentially reflecting adaptations to the shared environment and host tissue tropisms within clades.

Experimentally, I cloned and transformed genes from *Nematocida parisii* that are orthologs to essential yeast genes, transforming them into corresponding hetKO and ts yeast strains. The *N. parisii* ortholog for essential yeast gene *MPE1, NEPG-01868,* failed to rescue a temperature-sensitive strain for *MPE1* at restrictive temperatures, suggesting a possible functional divergence or loss-of-function of the ortholog in *N. parisii*. Conversely, heterozygous knockout complementation assays for the *N. parisii* ortholog to yeast *PSA1, NEPG-02059*, failed due to using an inadequate yeast strain for my hetKO assays.

All-in-all, this study is the first to probe the functional consequences of protein truncation in this highly reduced group of parasites. This study provides further insight into the extent to which these parasites are host-dependent and reveals novel protein-level adaptations to their parasitic lifestyles.

Methods - computational:

Code for analyses was written in R (R Core Team, 2021) and Bash (GNU, 2007). Scripts were organized into a pipeline using Snakemake (Köster and Rahmann, 2012). Code for this study was deposited to https://github.com/Jason-B-Jiang/microsporidia-protein-evolution. The full codebase and data/results files were uploaded to https://drive.google.com/file/d/11NNgs3qHuwrv6L1cJMvXqL-vCXZ9agTA/view?usp=sharing as a compressed archive.

Finding single-copy orthogroups between microsporidia and yeast proteomes

OrthoFinder v2.5.4 (Emms and Kelly, 2019) was used to find single-copy orthologs between each sequenced microsporidia species and yea. BLAST v2.9.0 (McGinnis and Madden, 2004) was used for the all-versus-all search step. The *S. cerevisiae* proteome was downloaded as a fasta file from Uniprot on January 11, 2022 (https://www.uniprot.org/proteomes/UP000002311). Proteomes for sequenced microsporidia species were obtained from Lina Wadi. The proteome for the *Microsporidia* outgroup species, *Rozella allomycis*, was also included. *Microsporidia* - yeast ortholog pairs were annotated as essential or non-essential using a list of essential yeast genes from Kofoed et al. (Kofoed et al., 2015, supplementary table 6).

For *Nematocida parisii*, OrthoFinder was run with proteomes of all sequenced *N. parisii* strains and yeast, following a similar approach as above. Proteomes for *Nematocida parisii* strains (1248, 1762, 2106, 2132, BG33, BG64, ERTm1, ERTm3) were downloaded from NCBI. Single-copy orthogroups for *N. parisii* were defined as single-copy orthologs conserved in at least two *N. parisii* strains and yeast.

These additional strains were included to maximize the number of orthogroups found, to have as many ortholog pairs as possible to consider for functional replacement experiments in yeast.

Assigning Pfam domains to microsporidia - yeast ortholog pairs

Protein domains were assigned to each ortholog using hmmscan from HMMER v3.3.2 (Finn et al., 2011) and Pfam 34.0 (Mistry et al., 2021), using Pfam domain family-specific bit-score cutoffs to determine the significance of domain assignments. Family-specific bit-score cutoffs were preferred over fixed E-value thresholds due to more accurate domain assignments with family-specific cutoffs (Punta et al., 2012). To obtain the final domain architecture of each ortholog, cath-resolve-hits v0.16.10 (Lewis et al., 2019) was run on the hmmscan outputs to resolve overlapping significant domain assignments. Only ortholog pairs with domains present in both of their final resolved domain architectures were retained for downstream analyses.

Comparing ortholog protein, domain and linker lengths

To compare overall length differences between Microsporidia - yeast ortholog pairs, a paired Mann-Whitney U test was performed on all pairs of orthologous protein lengths. To compare domain lengths between ortholog pairs, conserved domains in both orthologs were matched to each other using output domain architecture files from cath-resolve-hits. Domains occurring multiple times in an ortholog had their lengths combined into a single length for that domain. A paired two-sided Mann-Whitney U test was performed on all matched domains between ortholog pairs, with multiple matched domains possible from each ortholog pair. Linker length for each protein was calculated as $length\ of\ protein\ -\ length\ of\ all\ domains$. Significance for different in linker lengths between all ortholog pairs were also calculated using a paired two-sided Mann-Whitney U test.

Aligning single-copy ortholog domain architectures

Domain architectures for single-copy ortholog pairs were aligned as had been previously done (Forslund et al., 2011). Domains in domain architectures were first mapped back to their corresponding Pfam clans, as domains in the same clan are functionally and structurally conserved (Forslund et al., 2011; Mistry et al., 2021). Domain architectures for each ortholog were represented as strings of letters, with unique letters for each distinct Pfam clan. These domain architecture strings were aligned with the Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970), using the 'NameNeedle' package in R (Coombes, 2020). Matches were given a score of 0, gaps were given a score of -3, and mismatches were given a score of -10, as had been done previously (Forslund et al., 2011).

Identifying domain architectural changes between ortholog pairs

Three classes of domain architectural changes were defined: domain losses, domain gains, and domain swaps in the *Microsporidia* ortholog relative to its yeast ortholog. Domain losses in *Microsporidia* orthologs were inferred as a gap in the alignment of the *Microsporidia* domain architecture to its yeast ortholog. Domain gains were inferred as gaps in the aligned yeast ortholog domain architecture with its *Microsporidia* ortholog. Domain swaps were inferred as mismatching positions between the aligned domain architectures.

Calculating percent identities between ortholog pairs

Single-copy ortholog pairs were first aligned with the Needleman-Wunsch algorithm, using the 'needle' program from the EMBOSS software package (Rice et al., 2000). Alignment was performed with default scoring parameters and the BLOSUM45 substitution matrix. Percent identity between each aligned ortholog pair was calculated as $100 \times \frac{identical\ aligned\ residues}{aligned\ residues + alignment\ gaps}$.

Comparing extent of C-terminus truncation in *Microsporidia* to premature stop codon tolerance in yeast orthologs

To infer the extent of C-terminal truncation in each *Microsporidia* ortholog relative to its yeast ortholog, ortholog pairs were aligned as described above. C-terminal truncations were calculated as the length of the gap between the end of the yeast ortholog with the last aligned residue in the microsporidia ortholog. C-terminal truncations were normalized against gene lengths by dividing them by the length of their yeast orthologs. If the *Microsporidia* ortholog had a longer C-terminus than the yeast ortholog, then the C-terminal truncation value was positive instead of negative, representing a C-terminal truncation in the yeast ortholog relative to *Microsporidia.*

The number of premature stop codons (PTCs) tolerated by essential yeast genes was taken from Supplementary Table 6 from Sadhu et al.'s paper (Sadhu et al., 2018), from the 'HMM count of PTCs tolerated per gene' column. Only *Microsporidia* orthologs whose yeast orthologs were listed in this dataset were considered, and were annotated with the number of PTCs their yeast ortholog tolerated.

Identifying frequency of dispensable C-terminus domain loss in microsporidia orthologs

To infer dispensable C-terminal domains in essential yeast genes, I used previously published data on the tolerance of premature stop codons (PTCs) by essential yeast genes (Sadhu et al., 2018, supplemental table 11) . The 'dist_from_CDS_end' was taken to be the corresponding amino acid position in the protein sequence of the gene in which the stop codon was inserted. A dispensable C-terminal domain was a domain that could be disrupted by a PTC without impacting yeast viability. A PTC was considered to disrupt a protein domain if it occurs before the last residue of the coordinates for the domain.

To determine if dispensable C-terminal domains were more likely to be lost in *Microsporidia* orthologs to yeast, I only considered ortholog pairs with dispensable C-terminal domains in the yeast ortholog. *Microsporidia* orthologs were annotated with how many of their lost domains were dispensable C-terminal domains in the yeast ortholog. A one-proportion Z-test was used to determine the significance of the proportion of lost domains amongst all *Microsporidia* orthologs that have yeast orthologs with dispensable C-terminal domains.

Gene Ontology enrichment analysis of *Microsporidia* orthologs to yeast with lost protein domains

Gene Ontology enrichment analysis of *Microsporidia* orthologs with lost domains was performed with the PANTHER web service (http://www.pantherdb.org/) (Mi et al., 2021). A list of the UniProt names of yeast genes where the *Microsporidia* ortholog had undergone domain loss was submitted, using a background of Uniprot yeast gene names of all yeast genes conserved in *Microsporidia* as single-copy orthologs. The 'statistical overrepresentation' option was chosen, using the PANTHER GO slim collections of biological process, molecular function and protein class terms. The significance of enriched terms was determined with Fisher's exact test, at a False Discovery Rate threshold of <0.05.

Clustering yeast orthologs to Microsporidia by domain architecture conservation in Microsporidia orthologs

First, a table was constructed of all yeast genes conserved in >=1 *Microsporidia* species, and domain architecture conservation of each gene in its *Microsporidia* orthologs. Rows of the table were *Microsporidia* species and columns were yeast genes, with cells holding information on the domain architecture conservation of a particular *Microsporidia* - yeast ortholog.

With this table, a distance matrix was calculated for the yeast genes, based on their dissimilarity in domain architecture conservation across *Microsporidia* species. Gower's distance was used to calculate distances for these categorical variables, using the *daisy* function from the *cluster v2.1.2* R package (Maechler et al., 2021). Yeast genes were then clustered by similarities in domain architecture conservation patterns across *Microsporidia*, using the *agnes* function from the *cluster* R package for agglomerative hierarchical clustering (Murtagh and Legendre, 2014) with the distance matrix.

Clustering microsporidia species by lost Pfam clans

First, a list of all unique lost Pfam domains was collected for each *Microsporidia* species, using the aligned ortholog domain architectures for each species. The unique lost domains were then mapped back to their Pfam clans with ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.clans.tsv.gz, and further filtered to non-redundant clans only. If a domain did not belong to any clan, then the domain was kept as is. Pfam clans were used as domains in a clan have shared structure and function, thereby improving sensitivity in picking up domain loss patterns across species.

A dissimilarity matrix was constructed between each species and their set of unique clan losses using Gower's distance, with the *daisy* function from R package *cluster v2.1.2* (Maechler et al., 2021). Using this matrix, species were then clustered by their similiarity in lost clans using k-medoids clustering (Kaufman and Rousseeuw, 1990), using the *pam* function from *cluster*. The optimal number of clusters (k) for k-medoids clustering was selected by clustering with k = 2 to k = 30, and selecting the k-value that maximizes silhouette width (Rousseeuw, 1987) of the clustering, which is a measure of how well items fit in their clusters.

The clustered species were visualized using t-SNE (van der Maaten and Hinton, 2008), passing the dissimilarity matrix into the *Rtsne* function from the *Rtsne v0.15* R package (Krijthe, 2018). Maximum iterations for t-SNE was set to a high value of 5000, to ensure convergence to stable t-SNE plots. Ideal perplexity was chosen by iterating from 5 to $\frac{number\ of\ species - 1}{3}$, and manually inspecting the resulting t-SNE graphs. This range of perplexities was chosen based on the recommended range of perplexities of 5 - 50 from the original t-SNE paper (van der Maaten and Hinton, 2008), and from the documentation of *Rtsne* stating a maximum perplexity of $\frac{matrix\ rows - 1}{3}$.

Methods - experimental

PCR amplification of orthologous *Nematocida parisii* and *Saccharomyces cerevisiae* genes

Genes from *Nematocida parisii* and *Saccharomyces cerevisiae* were PCR amplified from *N. parisii ERTm1* and *S. cerevisiae BY4741* genomic DNA respectively, following manufacturer's instructions from Phusion. A two-step PCR was performed to introduce attB sites for the initial cloning of the PCR amplicons into entry vectors for Gateway cloning (Reece-Hoyes and Walhout, 2018). Primers for the first step were designed to amplify each gene without the stop codon, with partial attB sequences attached to the forward and reverse primers. Partial attB sequences were 5'-ACAAAAAAGCAGGCTCA-3' for forward primers and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTT-3' for reverse primers. Forward and reverse primers for amplifying all genes in this study are attached ('Primer sequences'). Primers for the second step were partial attB sequences that were complementary to the partial attB sites on the first primers, allowing amplification of the gene with the attB sequences. Amplicons from the two-step PCR were PEG-purified following instructions from Invitrogen, and successful amplification of genes was verified by agarose gel electrophoresis.

Gateway cloning of amplified orthologs into entry and expression vectors

PCR products from the two-step PCR were cloned with Gateway cloning to obtain expression clones. The PEG-purified PCR products were first inserted into the pDONR221 entry vector from Invitrogen by Gateway BP reaction, and transformed into competent E. coli cells to obtain entry clones. The entry clones were extracted from the cells using a Qiagen Miniprep kit, following the manufacturer's instructions. The extracted entry clones were then inserted into the pAG416-GPD-ccdB+6Stop destination vector (Laurent et al., 2020) by Gateway LR reaction, and transformed into competent E. coli to obtain expression clones. The expression clones were extracted as well to obtain the final expression clones. All entry and expression clones were verified by restriction digests and sequencing.

Transformation of expression clones into knockout yeast strains

Temperature-sensitive and hetKO yeast strains were obtained from Dr. Brenda Andrews. Transformations were done according to established methods (Gietz and Schiestl, 2007), using a short (<15 min) heat shock for temperature-sensitive strains. Expression clones for cloned yeast and *N. parisii* orthologs were transformed into their corresponding temperature-sensitive and hetKO yeast strains. The empty destination vector, pAG416-GPD-ccdB+6Stop, was also transformed into cells as a negative

control. Transformed temperature-sensitive cells were plated on -ura dextrose plates at 25 degC for 2 - 3 days, to select for successful transformants due to uracil metabolism deficiency in knockout strains and a ura3 cassette in the destination vector. Transformed hetKO cells were plated on -ura dextrose + G418 plates at 30 degC for 2 - 3 days, to select for both successful transformation and presence of G418-resistance knockout cassette.

Functional replacement assays in temperature-sensitive strains

Transformed temperature-sensitive cells were taken from their -ura dextrose plates and incubated in liquid -ura dextrose media, shaken at 200 rpm and 25 degC grown overnight. Three different overnight cultures were started for each transformant, as three independent biological replicates. The cultured cells were resuspended and replated on -ura dextrose plates using a 48-pronged spotter. The replated cells were incubated at 25 degC (permissive) or 37 degC (restrictive) for 48 hours, to test for rescue by the transformed expression clones. Cells transformed with the empty destination vector were not expected to grow at 37 degC. Growth of yeast colonies after 48 hours was quantified using ImageJ (Schneider et al., 2012), using established methods (Petropavlovskiy et al., 2020). ANOVA with Tukey posthoc correction was used to compare growth between different transformants.

Heterozygous knockout strain complementation testing

Transformed hetKO yeast cells were tested for complementation by yeast and *N. parisii* ortholog following previously established methods (Laurent et al., 2020). Single colonies of transformed cells were grown in 1 mL liquid -ura dextrose + G418 overnight, with vigorous shaking at 200 rpm. Three overnight cultures were started for each hetKO transformant, for three biological replicates under each condition. Cultured cells were replated on GNA-rich pre-sporulation plates for 24 hours, and multiple colonies from each plate were suspended in 1 mL sporulation mix (0.1% potassium acetate, 0.005% zinc acetate) until the mixture was slightly cloudy. The sporulation mix was incubated at 25 degC with vigorous shaking (200 rpm) for 3 - 5 days, and sporulation efficiency was assessed by microscopy. Sporulated cells were spun down and resuspended in water, and replated on Magic Marker plates (Laurent et al., 2020), with and without G418. Magic marker plates without G418 assessed overall sporulation efficiency, and plates with G418 tested for successful functional replacement of the gene knockout cassette in haploid cells by transformed plasmids (Laurent et al., 2020).

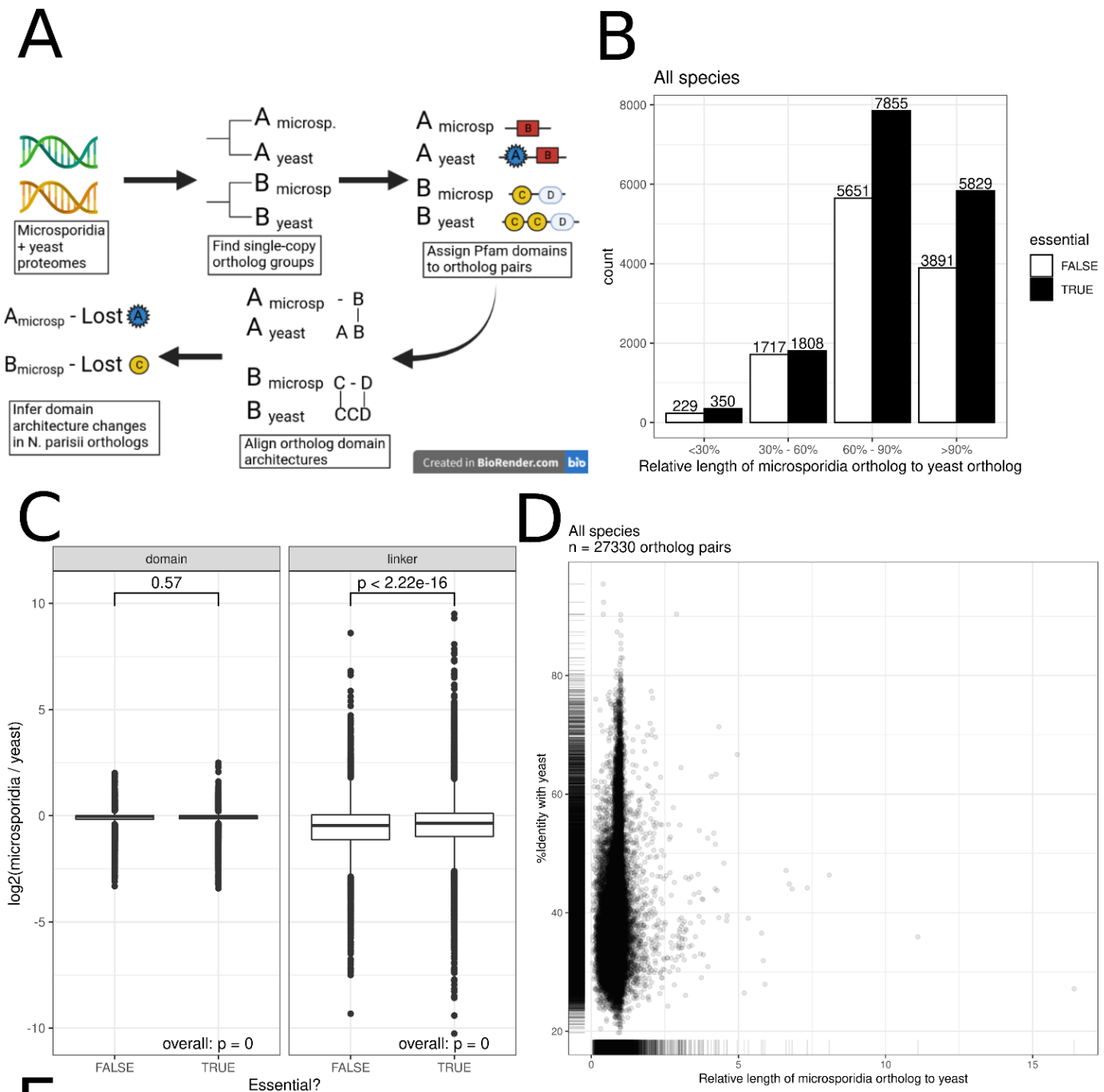

A
Microsporidia + yeast proteomes
Find single-copy ortholog groups
Assign Pfam domains to ortholog pairs
Align ortholog domain architectures
Infer domain architecture changes in N. parisii orthologs
Created in BioRender.com
B
All species
count
essential
FALSE
TRUE
229
350
1717
1808
5651
7855
3891
5829
<30%
30% - 60%
60% - 90%
>90%
Relative length of microsporidia ortholog to yeast ortholog
C
domain
linker
0.57
p < 2.22e-16
log2(microsporidia / yeast)
overall: p = 0
overall: p = 0
FALSE
TRUE
FALSE
TRUE
Essential?
D
All species
n = 27330 ortholog pairs
%Identity with yeast
Relative length of microsporidia ortholog to yeast
E
All microsporidia
15067 ortholog pairs
spearman rho = 0.26, p = 0
C-terminus difference / yeast ortholog length
Number of PTCs tolerated by yeast ortholog

FIG 1 *Microsporidia* proteins are reduced and divergent relative to their yeast orthologs **(A)** Computational workflow for identifying and comparing *Microsporidia* single-copy orthologs to yeast. **(B)** Relative lengths of *Microsporidia* proteins to their yeast orthologs, with relative length being $\frac{length\ of\ Microsporidia\ protein}{length\ of\ yeast\ ortholog}$. Each bin is labelled with the number of ortholog pairs falling into the bin. **(C)** Relative lengths of *Microsporidia* protein domains and linkers to their yeast orthologs, across yeast-essential and non-essential ortholog pairs. Relative *Microsporidia* domain/linker lengths are expressed as $log_2(\frac{length\ of\ Microsporidia\ domain/linker}{length\ of\ yeast\ ortholog\ domain/linker})$. P-values for overall difference between *Microsporidia* and yeast domain/linker lengths are at the bottom of each boxplot (see methods). P-values for difference in relative domain/linker lengths between essential and non-essential ortholog pairs were calculated with an unpaired Mann-Whitney U test. P-values for overall difference between *Microsporidia* and yeast domain/linker lengths were calculated with paired Mann-Whitney U tests. **(D)** Relative lengths of *Microsporidia* - yeast orthologs and their protein sequence identities (see methods). **(E)** Extent of C-terminus truncation in *Microsporidia* orthologs to yeast and premature stop codons tolerated by the yeast ortholog (see methods). Correlation between C-terminus truncation and number of tolerated PTCs was calculated with Spearman rank correlation.

Results

Microsporidia share an essential, reduced and diverged set of single-copy orthologs with *Saccharomyces cerevisiae*

To discover single-copy orthologs between *Microsporidia* species and yeast, I ran OrthoFinder (Emms and Kelly, 2019) on the proteomes of 46 *Microsporidia* species with the yeast proteome (Fig. 1A, methods). I considered single-copy orthologs, as these are more likely to be functionally conserved to their yeast orthologs. Paralogous gene relationships can have some paralogs diverge in function (Forslund et al., 2011; Soria et al., 2014).

Each species shared 599 single-copy orthologs on average with yeast (Supplementary Table S1), with ~58% of these being orthologous to essential yeast genes. These conserved orthologs were ~17.2% shorter than their yeast counterparts (Supplementary Table S1, Fig 1B), having both shorter linkers and domains (Fig 1C). The most reduction in these orthologs occurred in linkers (Supplementary Table S2) as opposed to domains (Supplementary Table S3), with *Microsporidian* linkers being 84% of their corresponding yeast ortholog linker lengths and *Microsporidian* protein domains being ~97.3% of their corresponding yeast ortholog domain lengths. Moreover, linkers in *Microsporidia* orthologs to essential yeast genes were less reduced (Fig 1C). *Microsporidia* orthologs were also highly diverged in sequence content from their yeast orthologs, sharing only ~39.4% sequence identity on average (Supplementary Table S4). Furthermore, sequence identity was poorly correlated with the relative lengths of *Microsporidia* proteins to their yeast orthologs (Fig 1D).

A previous study assessed the dispensability of C-terminal regions in essential yeast proteins, using CRISPR-Cas9 to insert premature stop codons (PTCs) in essential yeast genes (Sadhu et al., 2018). With this study, I investigated whether essential yeast genes that are more PTC-tolerant have *Microsporidia* orthologs with greater C-terminus truncation (see methods). Surprisingly, the number of PTCs tolerated by essential yeast genes was poorly correlated to the extent of C-terminus truncation in their *Microsporidia* orthologs (Fig 1E).

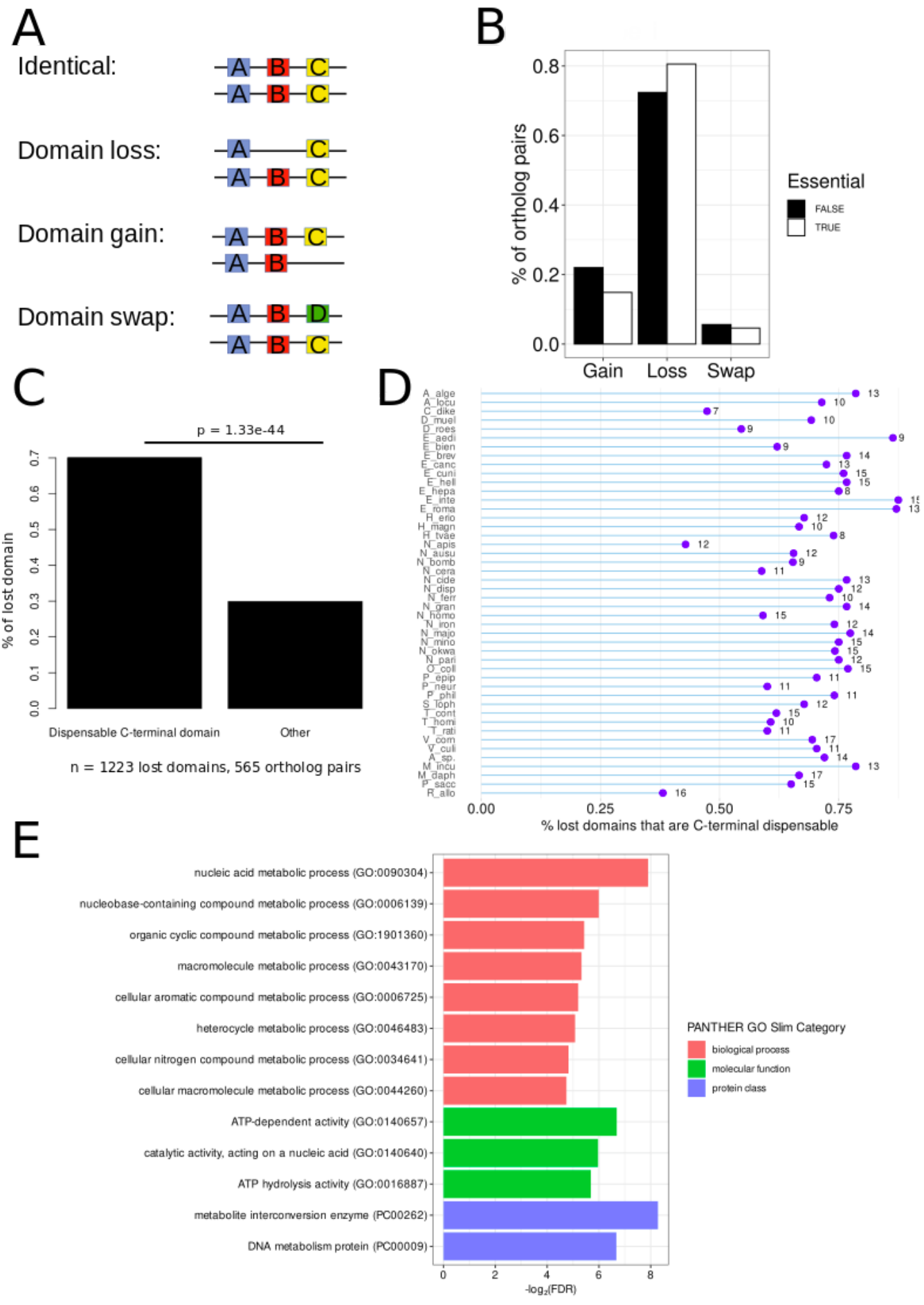

A
Identical:
A B C
A B C
Domain loss:
A C
A B C
Domain gain:
A B C
A B
Domain swap:
A B D
A B C
B
% of ortholog pairs
0.0
0.2
0.4
0.6
0.8
Gain
Loss
Swap
Essential
FALSE
TRUE
C
p = 1.33e-44
% of lost domain
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Dispensable C-terminal domain
Other
n = 1223 lost domains, 565 ortholog pairs
D
A_alge 13
A_locu 10
C_dike 7
D_muel 10
D_roes 9
E_aedi 9
E_bien 9
E_brev 14
E_canc 13
E_cuni 15
E_heli 15
E_hepa 8
E_inte 15
E_roma 13
R_erio 12
H_magn 10
H_tvae 8
N_apis 12
N_ausu 12
N_bomb 9
N_cera 11
N_cide 13
N_disp 12
N_ferr 10
N_gran 14
N_homo 15
N_iron 12
N_majo 14
N_mino 15
N_okwa 15
N_pari 12
O_coll 15
P_epip 11
P_neur 11
P_phil 11
S_loph 12
T_cont 15
T_homi 10
T_rati 11
V_corn 17
V_culi 11
A_sp. 14
M_incu 13
M_daph 17
P_sacc 15
R_allo 16
0.00
0.25
0.50
0.75
% lost domains that are C-terminal dispensable
E
nucleic acid metabolic process (GO:0090304)
nucleobase-containing compound metabolic process (GO:0006139)
organic cyclic compound metabolic process (GO:1901360)
macromolecule metabolic process (GO:0043170)
cellular aromatic compound metabolic process (GO:0006725)
heterocycle metabolic process (GO:0046483)
cellular nitrogen compound metabolic process (GO:0034641)
cellular macromolecule metabolic process (GO:0044260)
ATP-dependent activity (GO:0140657)
catalytic activity, acting on a nucleic acid (GO:0140640)
ATP hydrolysis activity (GO:0016887)
metabolite interconversion enzyme (PC00262)
DNA metabolism protein (PC00009)
0
2
4
6
8
-log2(FDR)
PANTHER GO Slim Category
biological process
molecular function
protein class

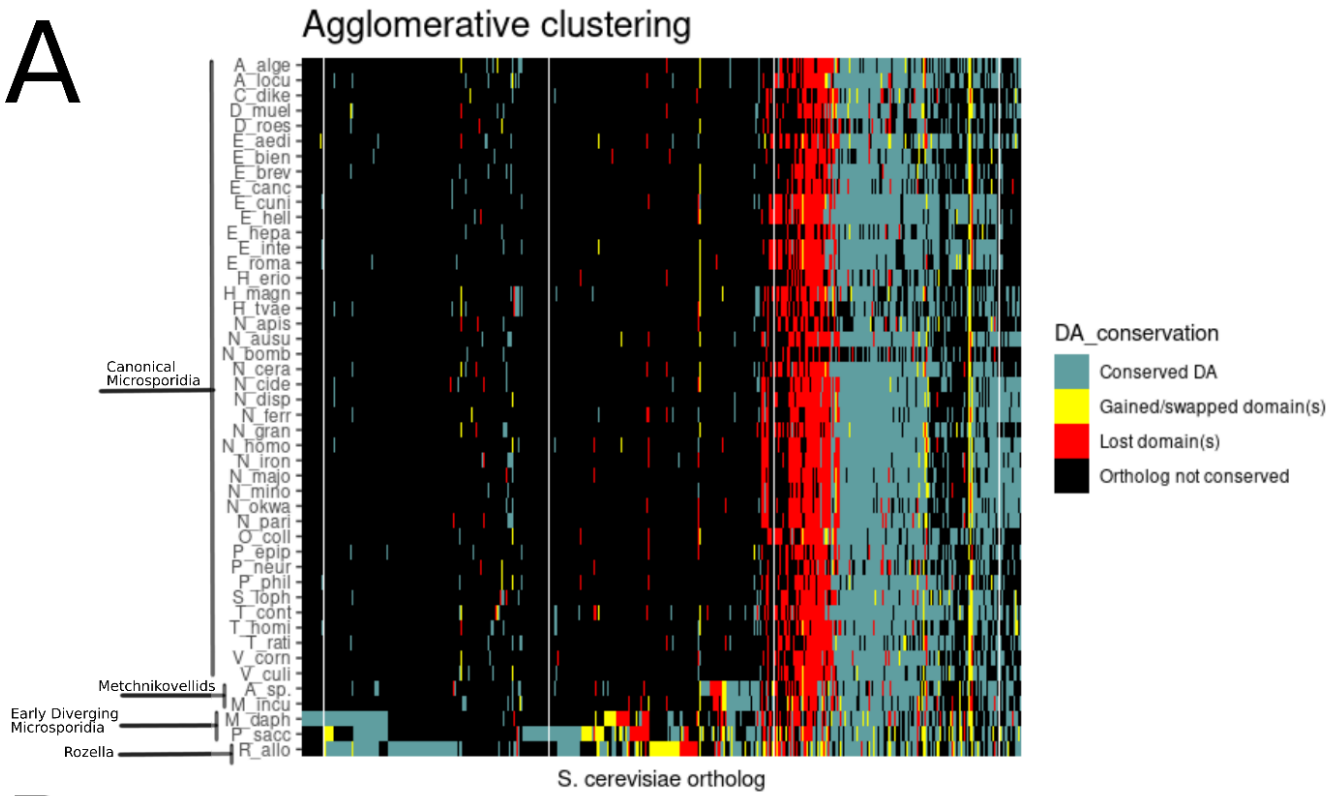
FIG 2 Domain loss is the most prevalent domain architectural change in *Microsporidia* orthologs to yeast. **(A)** Illustrations of domain architectural change events between *Microsporidia* and yeast orthologs. **(B)** Rates of domain architectural change events in *Microsporidia* orthologs to yeast, amongst *Microsporidia* orthologs with non-conserved domain architectures to yeast. Rates shown are relative to essential and non-essential ortholog pairs. **(C)** Proportion of lost domains in *Microsporidia* orthologs to essential yeast genes with dispensable C-terminal domains, that are these dispensable domains. P-value of this proportion was calculated with a one-proportion Z-test (see methods). **(D)** Rates of dispensable C-terminal domain loss across *Microsporidia* species. Points are labelled by the number of orthologs a species has with dispensable C-terminal domains in yeast. **(E)** Enriched Gene Ontology (GO) terms amongst yeast orthologs to *Microsporidia*, where the *Microsporidia* ortholog has undergone domain loss. A False-Discovery rate of FDR < 0.05 was set for significantly enriched GO terms (see methods). All significantly enriched terms in each category are shown.

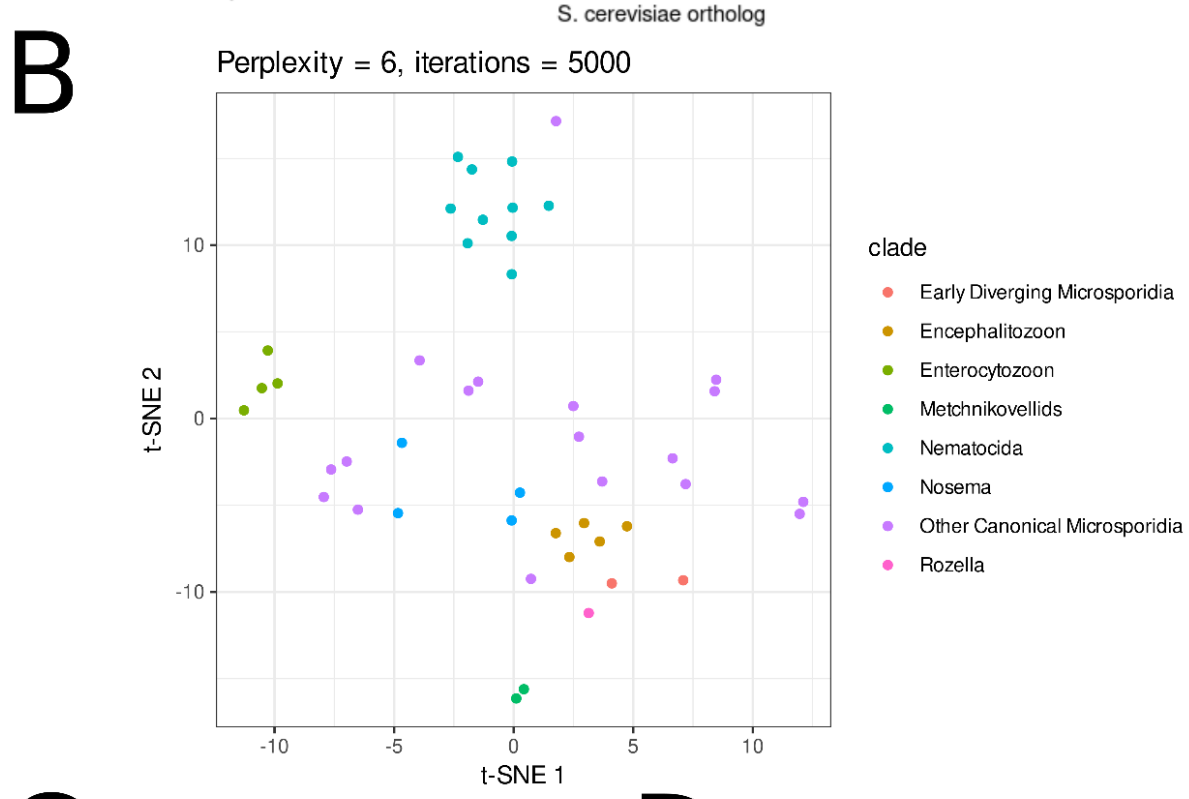
Domain architecture analysis of microsporidia single-copy orthologs to *Saccharomyces cerevisiae* reveals widespread domain loss

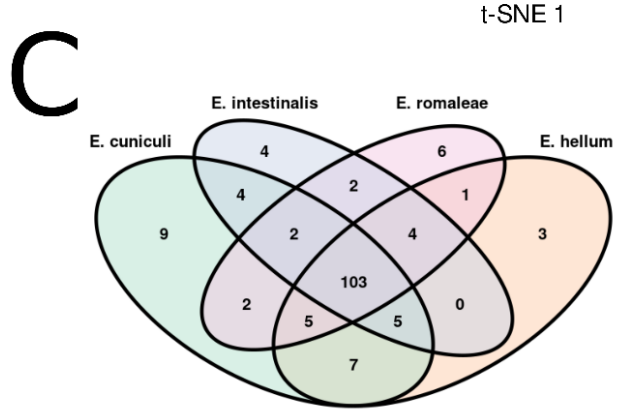
To better understand how protein reduction affected the structure of *Microsporidia* proteins, I expanded my bioinformatics pipeline to assign and compare domain architectures between *Microsporidia* and yeast single-copy orthologs (see methods). Three major domain architectural changes were identified in *Microsporidia* orthologs to yeast: domain loss, domain gain and domain swap (Fig 2A). Domain architecture between ortholog pairs was broadly conserved, with 68.5% of *Microsporidia* orthologs having identical domain architectures to their yeast orthologs (Supplementary Table S5). However, domain loss was the most prevalent architectural change, occurring in 25.4% of all ortholog pairs and comprising 80.7% of all domain architectural changes (Fig 2B, Supplementary Table S5). Intriguingly, rates of domain loss were similar between essential and non-essential *Microsporidia* - yeast ortholog pairs (Fig 2B). Rates of overall domain architecture conservation were also similar between essential and non-essential ortholog pairs (70.6% vs 67.3% of orthologs with identical domain architectures) (Supplementary Table S5).

Next, I investigated whether these lost domains were more likely to be inessential to the functioning of their yeast ortholog. To do so, I constructed a list of dispensable C-terminal domains in essential yeast genes from a dataset of premature stop codon tolerance by essential yeast genes (Sadhu et al., 2018). Protein domains that could be disrupted by PTCs without affecting the survival of yeast were considered dispensable C-terminal domains (see methods). Using this list, I investigated whether *Microsporidia* orthologs to these yeast genes with dispensable domains were more likely to lose these domains over others. Indeed, 70% of all domain losses in these orthologs were these dispensable C-terminal domains (Fig 2C). Moreover, the rate of dispensable domain loss was uniformly high across *Microsporidia* species (>50% of lost domains amongst orthologs to yeast genes with dispensable domains) (Fig 2D).

To elucidate broader functional consequences of domain loss in *Microsporidia* proteins, I considered enriched Gene Ontology (GO) terms for *Microsporidia* orthologs that have undergone domain loss (Fig 2E) (see methods). The most significantly enriched biological process term was ‘nucleic acid metabolic process, with every other biological process term except ‘cellular macromolecule metabolic process’ being a child term to that. Moreover, two of the three enriched terms for molecular function were related to ATP usage (‘ATP-dependent activity’ and ‘ATP hydrolysis’).


A
Agglomerative clustering
A_alge
A_locu
C_dike
D_muel
D_roes
E_aedi
E_bien
E_brev
E_canc
E_cuni
E_hell
E_hepa
E_inte
E_roma
H_erio
H_magn
H_tvae
N_apis
N_ausu
N_bomb
N_cera
N_cide
N_disp
N_ferr
N_gran
N_homo
N_iron
N_majo
N_mino
N_okwa
N_pari
O_coll
P_epip
P_neur
P_phil
S_toph
T_cont
T_homi
T_rati
V_corn
V_culi
A_sp.
M_incu
M_daph
P_sacc
R_allo
Canonical Microsporidia
Metchnikovellids
Early Diverging Microsporidia
Rozella
S. cerevisiae ortholog
DA_conservation
Conserved DA
Gained/swapped domain(s)
Lost domain(s)
Ortholog not conserved


B
Perplexity = 6, iterations = 5000
t-SNE 2
t-SNE 1
10
0
-10
-10
-5
0
5
10
clade
Early Diverging Microsporidia
Encephalitozoon
Enterocytozoon
Metchnikovellids
Nematocida
Nosema
Other Canonical Microsporidia
Rozella


C
E. intestinalis
E. romaleae
E. cuniculi
E. hellum
9
4
4
2
6
1
2
4
3
103
2
5
5
0
7

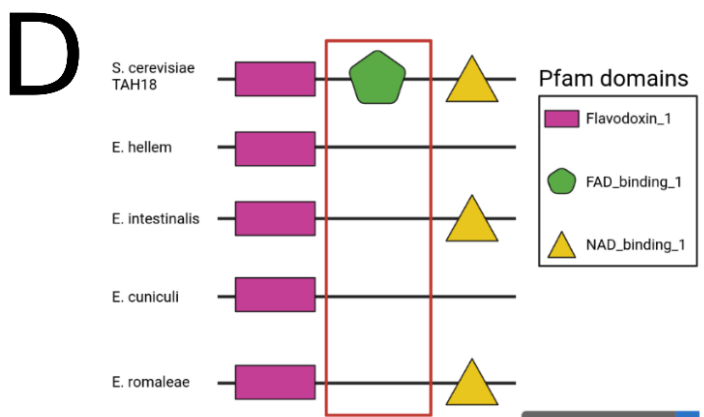

D
S. cerevisiae TAH18
E. hellem
E. intestinalis
E. cuniculi
E. romaleae
Pfam domains
Flavodoxin_1
FAD_binding_1
NAD_binding_1

FIG 3 Domain architectural changes in *Microsporidia* proteins relative to yeast occurred early in evolution, in patterns consistent with phylogeny. **(A)** Domain architecture conservation status of *Microsporidia* orthologs to yeast. Columns are yeast orthologs found in at least 1 microsporidia species, and rows are orthologs from a single microsporidia species. Yeast orthologs (columns) were clustered by their similarity in domain architecture conservation patterns across *Microsporidia* species (see methods). Species are arranged according to their broader phylogenetic clades. **(B)** t-SNE of *Microsporidia* species and their domain loss patterns (see methods). Points are coloured by their *Microsporidia* phylogenetic clades. **(C)** Overlap of lost Pfam clans in *Encephalitozoon* species. **(D)** Illustration of the conserved loss of *FAD_binding_1* domain in *Encephalitozoon spp.* orthologs to *TAH18*.

Domain architectural changes in *Microsporidia* orthologs to yeast occurred early in evolution in clade-specific patterns

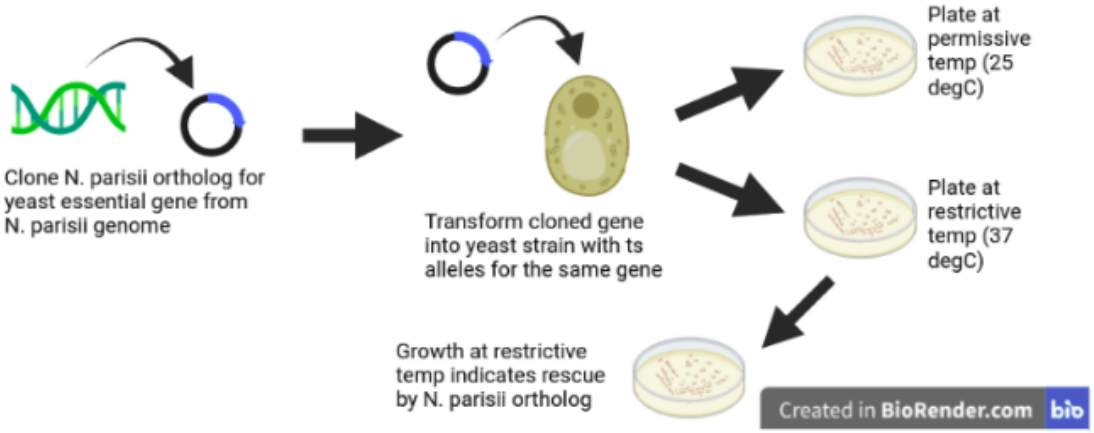
Next, I investigated whether domain architectural changes in *Microsporidia* orthologs to yeast followed any phylogenetic patterns. First, I clustered yeast orthologs to *Microsporidia* by the domain architectures of their *Microsporidia* orthologs (Fig 3A) (see methods). The clustering revealed a core set of yeast genes conserved across *Microsporidia*, with mostly conserved domain architectures. Moreover, there was a prominent cluster of genes that lost domains together in *Microsporidia*. Some domain architectural changes also seemed to occur early in evolution, with some genes losing domains as early as in the outgroup species *Rozella allomycis*. However, there were also incidences of individual species and clade-restricted domain losses and gains.

Given that groups of genes seemed to lose domains together in *Microsporidia*, I investigated whether there were any phylogenetic patterns to the lost domains themselves. I clustered *Microsporidia* species using lost domains in their yeast orthologs (Fig 3B) (see methods). Indeed, *Microsporidia* have undergone patterns of domain loss consistent with their phylogenetic groupings (Fig 3B, Supplementary Table S6). For instance, species were separated into their broader evolutionary clades, with *Canonical Microsporidia*, *Metchnikovellids*, *Early Diverging Microsporidia* and outgroup species *Rozella allomycis* falling into distinct clusters. Moreover, I observed genera-specific patterns of domain loss, with *Nematocida* and *Encephalitozoon* falling in their own distinct clusters.

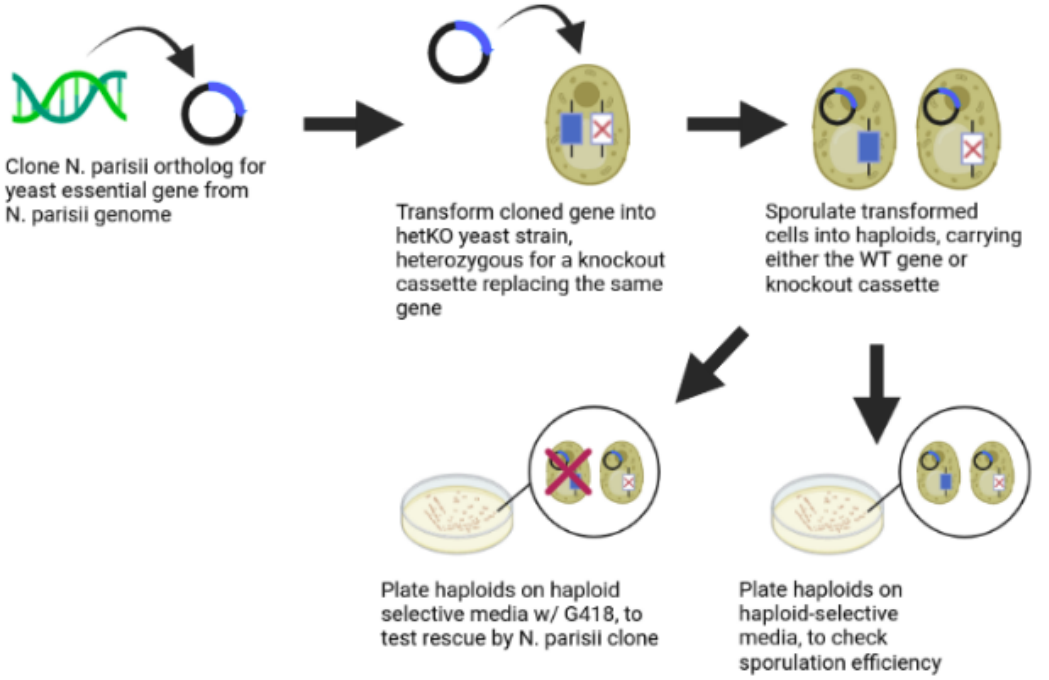
To further elucidate genera-specific patterns of domain loss, I investigated patterns of domain loss in *Encephalitozoon*. *Encephalitozoon* share a common core of lost domains belonging to 103 Pfam clans, wherein around three-quarters of lost domains in each *Encephalitozoon* species belong to these clans (Fig 3C). An example of such a lost Pfam clan was the *FAD_Lum_binding* (Riboflavin synthase/Ferredoxin reductase FAD-binding domain) clan of domains, encompassing domains binding essential cofactors for redox reactions (Dym and Eisenberg, 2001). Further investigation of

FAD_Lum_binding losses in *Encephalitozoon* revealed that all the losses occurred in orthologs to yeast *TAH18*. *TAH18* is an essential mitochondrial gene for cytosolic Iron-Sulfur cluster assembly (CIA) (Saccharomyces Genome Database, 2022), using FAD to transfer electrons from NADPH to *DRE2*, another essential component for CIA (Netz et al., 2010; Soler et al., 2011). In particular, the *FAD_binding_1* domain from the *FAD_Lum_binding* clan was uniformly lost in all *Encephalitozoon* orthologs to *TAH18* (Fig 3D).

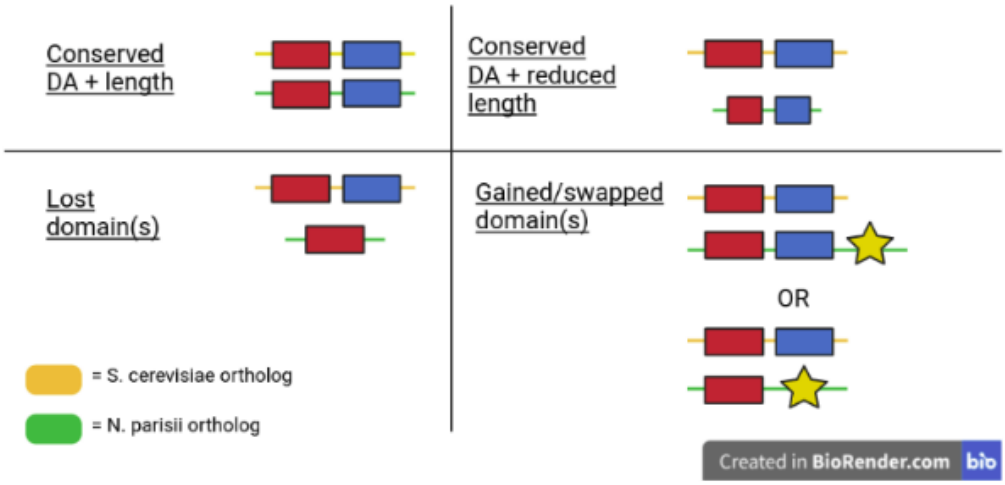
A



B



C



D

Yeast gene	*N. parisii* ortholog	Yeast gene length (AA)	*N. parisii* ortholog length (AA)	Ortholog category
MPE1 (P35728)	NEPG-01868	263	441	Gained domain(s)
PSA1 (P41940)	NEPG-02059	363	361	Conserved domain architecture + length

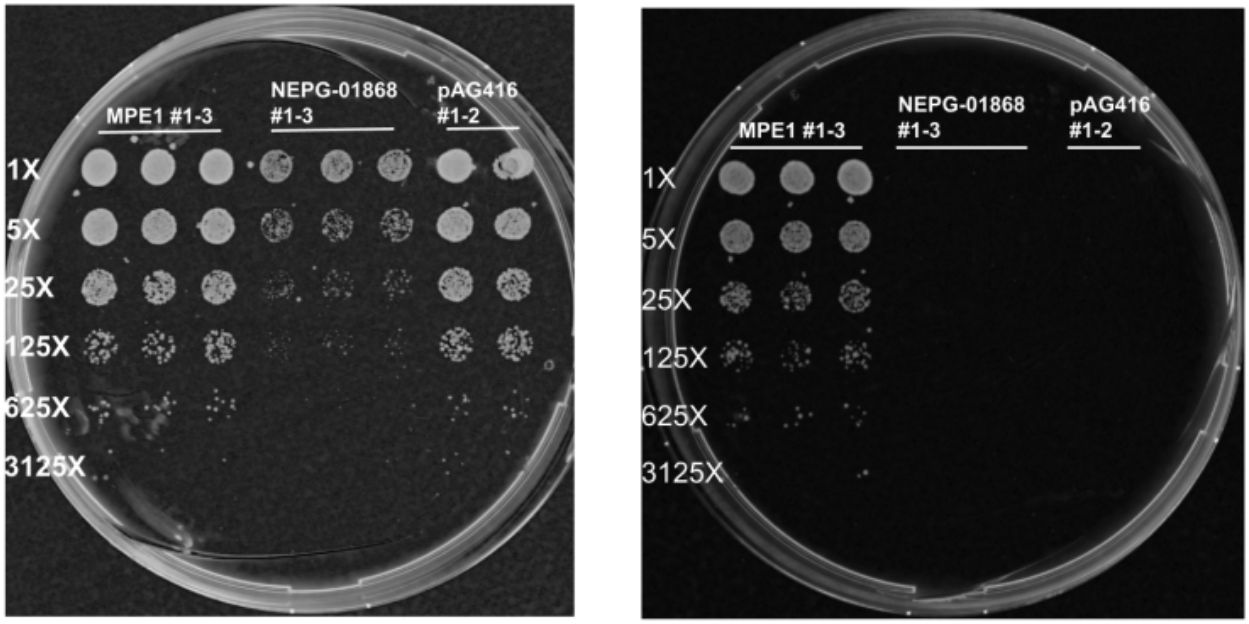
FIG 4 In-vivo gene complementation assays of essential *Nematocida parisii* orthologs in yeast. **(A)** Experimental workflow for gene complementation assays in temperature-sensitive yeast strains for essential genes. **(B)** Experimental workflow for gene complementation assays in *hetKO* yeast strains for essential genes. **(C)** Testing categories for *N. parisii* - yeast single-copy ortholog pairs chosen for complementation **(D)** Table of microsporidia - yeast ortholog pairs tested to date **(D)** Quantitative analysis of dilution assays for ts-mpe1 **(E)** Lengths and testing categories for *MPE1* and *NEPG-01868*

Designing functional replacement assays for *Nematocidia parisii* orthologs in yeast

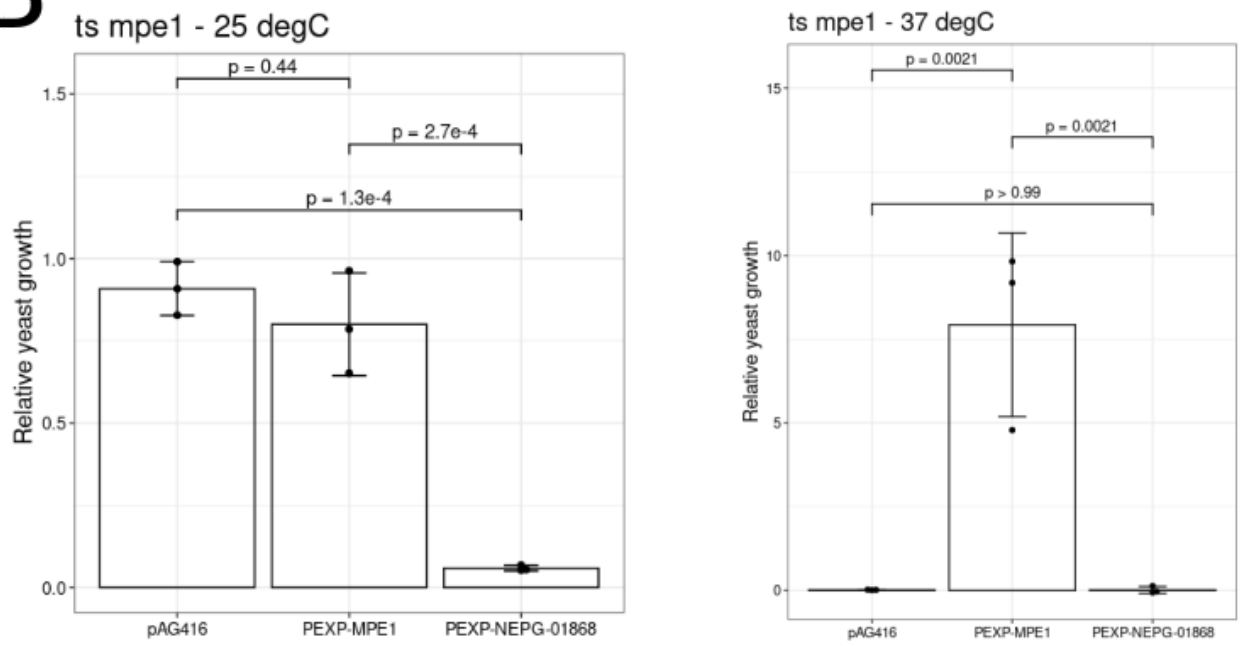
To test the functional conservation of *Microsporidia* genes to their yeast orthologs, I cloned single-copy orthologs to essential yeast genes from *Nematocida parisii*. I transformed the cloned *N. parisii* orthologs into corresponding temperature-sensitive (ts) and heterozygous knockout (hetKO) yeast strains, to see if the orthologs can rescue viability in these strains (Fig 4A, 4B). The corresponding orthologs from yeast were also cloned and transformed as a positive control in these functional replacement assays. *N. parisii* was chosen as a "representative" *Microsporidia* for these functional assays, as it has been well studied as a natural pathogen of the model organism *Caenorhabditis elegans* (Luallen et al., 2016).

To select *N. parisii* - yeast single-copy ortholog pairs for these functional assays, I adjusted my bioinformatics pipeline to consider all sequenced *N. parisii* strains for finding orthogroups between *N. parisii* and yeast (see methods). This was done to increase the number of single-copy orthogroups detected, maximizing the number of potential ortholog pairs to test. To select ortholog pairs for functional replacement assays, orthogroups were filtered to only those for essential yeast genes and by the availability of ts and hetKO strains. Moreover, orthogroups were filtered by the absence of insignificant hmmscan hits in the *N. parisii* ortholog for putatively missing domains, to minimize the possibility of testing false-positive domain loss ortholog pairs (see methods). A total of 632 single-copy orthogroups were discovered between *N. parisii* and yeast, with 289 single-copy orthogroups selected for the replacement assays (Supplemental Table S6). The 289 testable ortholog pairs were further divided into four categories, based on relative lengths and domain architectures of the *N. parisii* ortholog to their yeast orthologs (Fig 4C, Supplemental Table S6). Orthologs were considered to have similar lengths if the *N. parisii* ortholog was within 15% of its yeast ortholog's length. This number was arbitrarily chosen, based on the observation that *Encephalitozoon cuniculi* proteins are ~15% shorter on average than their yeast orthologs (Katinka et al., 2001). For this study, I selected *N. parisii* orthologs to yeast *MPE1* and *PSA1*, *NEPG-01868* and *NEPG-02059* respectively, for these functional complementation experiments in yeast (Fig 4D).

A


MPE1 #1-3
NEPG-01868 #1-3
pAG416 #1-2
1X
5X
25X
125X
625X
3125X
MPE1 #1-3
NEPG-01868 #1-3
pAG416 #1-2
1X
5X
25X
125X
625X
3125X

B


ts mpe1 - 25 degC
p = 0.44
p = 2.7e-4
p = 1.3e-4
Relative yeast growth
1.5
1.0
0.5
0.0
pAG416
PEXP-MPE1
PEXP-NEPG-01868
ts mpe1 - 37 degC
p = 0.0021
p = 0.0021
p > 0.99
Relative yeast growth
15
10
5
0
pAG416
PEXP-MPE1
PEXP-NEPG-01868

C

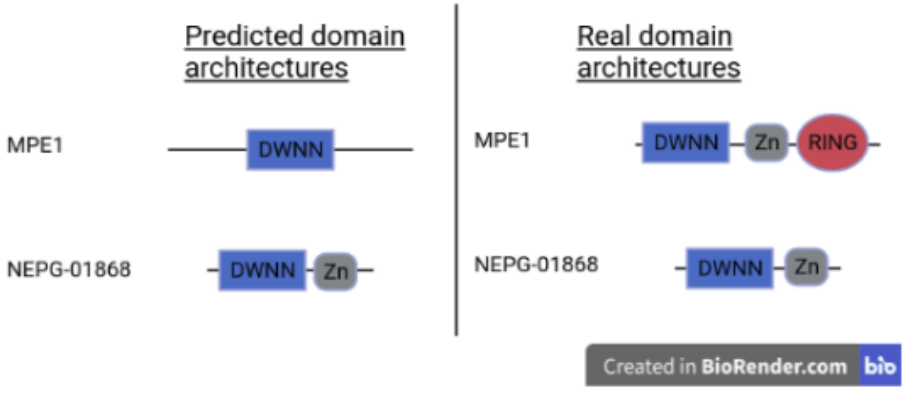

Predicted domain architectures
MPE1
DWNN
NEPG-01868
DWNN
Zn
Real domain architectures
MPE1
DWNN
Zn
RING
NEPG-01868
DWNN
Zn
Created in BioRender.com
bio

FIG 5 N. parisii NEPG-01868 fails to complement MPE1 in a temperature-sensitive strain for MPE1. **(A)** Spotting assays for ts-mpe1 transformed with clones for MPE1, NEPG-01868 and the empty clone vector (pAG416GPD-ccdB). Three biological replicates were plated for each condition, with a 5X serial dilution down plate rows. The third biological replicate for pAG416 was plated on a separate plate for each temperature (not shown) due to lack of plate space. **(B)** Comparison of mean growth in transformants under the permissive and restrictive temperatures, using colonies from the 4th dilution (125X) for each plate. Colony growth was obtained by measuring mean gray values of spots from the plate images in (A), and mean gray values for each condition were divided by the mean gray value of pAG416 transformants. Statistical significance of differences in mean growth between conditions were determined using one-way ANOVA and Tukey p-value correction for multiple testing. **(C)** Comparison of computationally predicted domain architectures for MPE1 and NEPG-01868, and the most probable domain architectures from literature.

The *Nematocida parisii* ortholog to yeast *MPE1*, *NEPG-01868*, fails to rescue a temperature-sensitive MPE1 strain

So far, I have performed functional complementation assays for *MPE1* in a temperature-sensitive *MPE1* strain, and both *MPE1* and *PSA1* in heterozygous knockout strains. For functional complementation assays in temperature-sensitive *MPE1*, *ts-mpe1* cells were transformed with cloned *NEPG-01868*, as well as cloned *MPE1* from yeast as a positive control and the empty destination vector (pAG416GPD-ccdB) as a negative control (see methods). The transformed cells were then spotted onto -ura dextrose plates to select for successfully transformed cells, and to quantify growth differences between different transformants (see methods) (Fig 5A, 5B). At the permissive temperature of 25 degC, I observed similar growth between *ts-mpe1* transformed with its cognate yeast gene and the empty destination vector (Fig 5A, 5B). However, growth in cells transformed with *NEPG-01868* was reduced relative to the two other transformants, suggesting toxicity from the expression of *NEPG-01868* in *ts-mpe1*. However, at a restrictive temperature of 37 degC, only *MPE1* itself could rescue *ts-mpe1*, with no growth from *NEPG-01868* or the empty vector (Fig 5A, 5B). Taken together, these results for *NEPG-01868* and *MPE1* in *ts-mpe1* suggest an inability for *NEPG-01868* to functionally replace its yeast ortholog *MPE1*.

Conversely, my functional complementation assays have been unsuccessful in my *MPE1* and *PSA1* hetKO strains. I have transformed the cloned *N. parisii* orthologs, yeast genes and empty destination vector into their corresponding hetKO strains. So far, no sporulated cells for any transformants have been able to grow on any of the Magic Marker plates used for haploid selection (see methods). My hetKO assays have likely failed because the hetKO strains used in this study are of a different genetic background from the hetKO strains used to develop Magic Marker selection (Laurent et al., 2020; personal communication with Dr. Aaron Reinke).

Discussion

Microsporidia have strategically reduced their proteins to retain function

Through my bioinformatics pipeline, I observe *Microsporidia* sharing a set of reduced and mostly essential orthologs to yeast, corroborating with a previous study reporting that *Microsporidia* share a common core of essential and highly expressed genes in yeast (Nakjang et al., 2013). This reduction may suggest the exploitation of a growth advantage to having shorter proteins (Wang et al., 2011), and a pressure to retain critical genes for survival while stripping away the rest. Moreover, linkers being more reduced than domains may reflect a pressure to conserve function in truncated proteins by retaining as many of the functional domains as possible. However, this observation may also be due to the limits of domain detection by Hidden Markov Models (HMMs) for protein domains. Highly reduced but functional domains in *Microsporidia* proteins may not be picked up by HMMs. Moreover, linkers being less reduced in essential than non-essential *Microsporidia* orthologs can also reflect a need to conserve protein-protein interactions in these essential genes, as linker regions often facilitate these interactions (Wang et al., 2011). Furthermore, the overall shortening of linkers in *Microsporidia* may reflect their reduced genomes and fewer protein interaction partners to maintain. The counterintuitive observation that *Microsporidia* do not retain more C-terminus sequence in orthologs to essential yeast genes that are less tolerant to PTCs may suggest a dispensability of these C-terminal features in the *Microsporidia* orthologs, or these genes no longer being essential to *Microsporidia*.

Prevalent domain loss in Microsporidia may reflect functionally conservative protein reduction and adaptation to host-dependent lifestyles

The prevalent domain loss observed in *Microsporidia* orthologs to yeast may have three explanations. First, *Microsporidia* may be stripping away dispensable protein domains to minimize protein length while conserving function. Indeed, I observed dispensable C-terminal domains being preferentially lost in *Microsporidia* when possible. However, this observation was limited to only *Microsporidia* orthologs with yeast genes that have known dispensable C-terminal domains, and only these C-terminal domains could be identified as dispensable with Sadhu et al's data. Second, due to the extensive host-dependency of *Microsporidia* (Wadi and Reinke, 2020), they may no longer require certain genes conserved with yeast, losing functional constraints to maintain domain architecture. GO enrichment analysis of orthologs that lost domains revealed an overrepresentation of nucleotide metabolism and ATP usage terms. This may reflect *Microsporidia*'s inability to make their own nucleotides (Dean et al., 2016) and reduced energy demands due to their host dependency. Thirdly, detection of protein domains by Hidden

Markov Models may be limited by the high sequence divergence in *Microsporidia* orthologs to yeast, potentially failing to detect domains.

Domain architectural changes in *Microsporidia* may reflect host and environmental adaptations

Microsporidia are highly specialized parasites, parasitizing a single or a narrow range of hosts in specific ecological environments (Murareanu et al., 2021). The broadly conserved patterns of domain architecture conservation and divergence I observed in *Microsporidia* orthologs to yeast may simply reflect the inheritance of domain conservation patterns from early ancestral species or earlier diverging species. However, more species and clade-specific instances of domain architecture conservation and divergence may reflect unique adaptations of these *Microsporidia* species to their hosts and environments. Similarly, the distinct patterns of lost domains across *Microsporidia* clades may simply reflect evolutionary similarity, but may also reflect specific adaptations of these *Microsporidia* to their lifestyles. Evolutionarily related *Microsporidia* were previously observed to share more common environments and host tissue tropisms (Murareanu et al., 2021). Thus, these specific patterns of domain loss may reflect the divergence of protein function to adapt or degradation of proteins that are no longer required for these shared conditions within clades.

Loss of the FAD_binding_1 domain in *Encephalitzoon* orthologs to yeast TAH18 suggests reduced dependence on cytosolic Iron-Sulfur clusters

Iron-Sulfur (Fe-S) clusters are typically assembled in the cytosol or mitochondria of eukaryotic cells, acting as essential factors for a broad range of proteins (Tsaousis, 2019). In yeast, Fe-S cluster synthesis is the only essential biosynthetic function of the mitochondria (Lill and Kispal, 2000; Tsaousis, 2019). Thus, the conserved loss of FAD_binding_1 in *Encephalitozoon* orthologs to the yeast mitochondrial gene TAH18 may have occurred as part of the degradation of *Microsporidian* mitochondria to simple mitosomes (Wadi and Reinke, 2020). Moreover, mitosomes in *Encephalitozoon cuniculi* were previously shown to have Fe-S cluster synthesis as a key function (Goldberg et al., 2008). Thus, this reduction of *TAH18* in *Encephalitozoon* may suggest an increased reliance on mitosomal Fe-S clusters, as opposed to cytosolic Fe-S clusters, and may serve as another example of the reduced metabolic capacity of *Microsporidia* (Keeling and Fast, 2002). However, the extent to which Fe-S cluster synthesis occurs in the cytosol over the mitosome in *Encephlitozoon,* or any other group of *Microsporidia,* has yet to be determined.

Nematocida parisii NEPG-01868 has diverged or lost function relative to its yeast ortholog, MPE1

To elucidate the reasons for NEPG-01868 being unable to complement MPE1, the function of MPE1 and the domain architectures of the orthologs were considered. In yeast, MPE1 is an essential subunit of the Cleavage/Polyadenylation Factor (CPF) for mRNA 3' end processing (Vo et al., 2001). From the computational pipeline, MPE1 was predicted to only have a ubiquitin-like domain (DWNN), while NEPG-01868 was predicted to also have a zinc knuckle domain (Fig 5C). However, the predicted domain architecture for MPE1 is incorrect, as past studies have experimentally determined three essential domains in MPE1 - a ubiquitin-like domain, zinc knuckle and RING finger (Vo et al., 2001; Lee and Moore, 2014). Thus, it is more plausible that NEPG-01868 has actually lost a RING finger relative to MPE1, especially considering its heavy truncation (263 aa vs 441 aa). The RING finger is required by MPE1 for RNA binding (Lee and Moore, 2014). The loss of this domain in NEPG-01868 may reflect N. parisii no longer requiring MPE1 in its CPF. Alternatively, N. parisii may have acquired a novel CPF subunit to compensate for RING, similar to how Encephalitozoon cuniculi evolved novel ribosomal proteins to compensate for missing structural rRNAs (Nicholson et al., 2021).

References

Wadi, L., and Reinke, A.W. (2020). Evolution of microsporidia: An extremely successful group of eukaryotic intracellular parasites. PLoS Pathog *16*, e1008276.

Murareanu, B.M., Sukhdeo, R., Qu, R., Jiang, J., and Reinke, A.W. (2021). Generation of a Microsporidia Species Attribute Database and Analysis of the Extensive Ecological and Phenotypic Diversity of Microsporidia. MBio *12*, e0149021.

Nakjang, S., Williams, T.A., Heinz, E., Watson, A.K., Foster, P.G., Sendra, K.M., Heaps, S.E., Hirt, R.P., and Martin Embley, T. (2013). Reduction and Expansion in Microsporidian Genome Evolution: New Insights from Comparative Genomics. Genome Biology and Evolution *5*, 2285–2303.

Keeling, P.J., and Fast, N.M. (2002). Microsporidia: Biology and Evolution of Highly Reduced Intracellular Parasites. Annu. Rev. Microbiol. *56*, 93–116.

Corradi, N., and Slamovits, C.H. (2011). The intriguing nature of microsporidian genomes. Brief Funct Genomics *10*, 115–124.

Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E.S., and Keeling, P.J. (2010). The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat Commun *1*, 77.

Keeling, P.J., Corradi, N., Morrison, H.G., Haag, K.L., Ebert, D., Weiss, L.M., Akiyoshi, D.E., and Tzipori, S. (2010). The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. Genome Biol Evol *2*, 304–309.

Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., et al. (2001). Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature *414*, 450–453.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research *49*, D412–D419.

Kachroo, A.H., Laurent, J.M., Yellman, C.M., Meyer, A.G., Wilke, C.O., and Marcotte, E.M. (2015). Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science *348*, 921–925.

Kachroo, A.H., Laurent, J.M., Akhmetov, A., Szilagyi-Jones, M., McWhite, C.D., Zhao, A., and Marcotte, E.M. (2017). Systematic bacterialization of yeast genes identifies a near-universally swappable pathway. Elife *6*, e25093.

Laurent, J.M., Garge, R.K., Teufel, A.I., Wilke, C.O., Kachroo, A.H., and Marcotte, E.M. (2020). Humanization of yeast genes with multiple human orthologs reveals functional divergence between paralogs. PLoS Biol *18*, e3000627.

Botstein, D., Chervitz, S.A., and Cherry, J.M. (1997). Yeast as a model organism. Science *277*, 1259–1260.

Li, Z., Vizeacoumar, F.J., Bahr, S., Li, J., Warringer, J., Vizeacoumar, F.S., Min, R., Vandersluis, B., Bellay, J., Devit, M., et al. (2011). Systematic exploration of essential yeast gene function with temperature-sensitive mutants. Nat Biotechnol *29*, 361–367.

Ben-Aroya, S., Coombes, C., Kwok, T., O'Donnell, K.A., Boeke, J.D., and Hieter, P. (2008). Toward a Comprehensive Temperature-Sensitive Mutant Repository of the Essential Genes of Saccharomyces cerevisiae. Molecular Cell *30*, 248–258.

Giaever, G., and Nislow, C. (2014). The yeast deletion collection: a decade of functional genomics. Genetics *197*, 451–465.

R Core Team (2021). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

GNU, P. (2007). Free Software Foundation. Bash (3.2. 48)[Unix shell program].

Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol *20*, 238.

McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res *32*, W20-25.

Kofoed, M., Milbury, K.L., Chiang, J.H., Sinha, S., Ben-Aroya, S., Giaever, G., Nislow, C., Hieter, P., and Stirling, P.C. (2015). An Updated Collection of Sequence Barcoded Temperature-Sensitive Alleles of Yeast Essential Genes. G3 (Bethesda) *5*, 1879–1887.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29-37.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Lewis, T.E., Sillitoe, I., and Lees, J.G. (2019). cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. Bioinformatics *35*, 1766–1767.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443–453.

Coombes, K.R. (2020). NameNeedle: Using Needleman-Wunsch to Match Sample Names.

Rice P, Longden I & Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277 Available at: https://ssbio.readthedocs.io/en/latest/instructions/emboss.html

Sadhu, M.J., Bloom, J.S., Day, L., Siegel, J.J., Kosuri, S., and Kruglyak, L. (2018). Highly parallel genome variant engineering with CRISPR-Cas9. Nat Genet *50*, 510–514.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., and Thomas, P.D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res *49*, D394–D403.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). cluster: Cluster Analysis Basics and Extensions.

Murtagh, F., and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif *31*, 274–295.

(1990). Partitioning Around Medoids (Program PAM). In Wiley Series in Probability and Statistics, L. Kaufman, and P.J. Rousseeuw, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 68–125.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics *20*, 53–65.

van der Maaten, L., and Hinton, G. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research *9*, 2579–2605.

Krijthe, J. (2018). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation.

Reece-Hoyes, J.S., and Walhout, A.J.M. (2018). Gateway Recombinational Cloning. Cold Spring Harb Protoc *2018*.

Gietz, R.D., and Schiestl, R.H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc *2*, 31–34.

Petropavlovskiy, A.A., Tauro, M.G., Lajoie, P., and Duennwald, M.L. (2020). A Quantitative Imaging-Based Protocol for Yeast Growth and Survival on Agar Plates. STAR Protocols *1*, 100182.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods *9*, 671–675.

Soria, P.S., McGary, K.L., and Rokas, A. (2014). Functional Divergence for Every Paralog. Molecular Biology and Evolution *31*, 984–992.

Dean, P., Hirt, R.P., and Embley, T.M. (2016). Microsporidia: Why Make Nucleotides if You Can Steal Them? PLoS Pathog *12*, e1005870.

Luallen, R.J., Reinke, A.W., Tong, L., Botts, M.R., Félix, M.-A., and Troemel, E.R. (2016). Discovery of a Natural Microsporidian Pathogen with a Broad Tissue Tropism in Caenorhabditis elegans. PLoS Pathog *12*, e1005724.

Vo, L.T., Minet, M., Schmitter, J.M., Lacroute, F., and Wyers, F. (2001). Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. Mol Cell Biol *21*, 8346–8356.

Lee, S.D., and Moore, C.L. (2014). Efficient mRNA polyadenylation requires a ubiquitin-like domain, a zinc knuckle, and a RING finger domain, all contained in the Mpe1 protein. Mol Cell Biol *34*, 3955–3967.

Nicholson, D., Salamina, M., Panek, J., Helena-Bueno, K., Brown, C.R., Hirt, R.P., Ranson, N.A., and Melnikov, S.V. (2021). "Lose-to-gain" adaptation to genome decay in the structure of the smallest eukaryotic ribosomes (Evolutionary Biology).

Dym, O., and Eisenberg, D. (2001). Sequence-structure analysis of FAD-containing proteins. Protein Sci *10*, 1712–1728.

(2022). TAH18 | SGD (Saccharomyces Genome Database). Available at https://www.yeastgenome.org/locus/S000006252

Tsaousis, A.D. (2019). On the Origin of Iron/Sulfur Cluster Biosynthesis in Eukaryotes. Front. Microbiol. *10*, 2478.

Goldberg, A.V., Molik, S., Tsaousis, A.D., Neumann, K., Kuhnke, G., Delbac, F., Vivares, C.P., Hirt, R.P., Lill, R., and Embley, T.M. (2008). Localization and functionality of microsporidian iron–sulphur cluster assembly proteins. Nature *452*, 624–628.

Keeling, P.J., and Fast, N.M. (2002). Microsporidia: biology and evolution of highly reduced intracellular parasites. Annu Rev Microbiol *56*, 93–116.

James, T.Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., and Stajich, J.E. (2013). Shared Signatures of Parasitism and Phylogenomics Unite Cryptomycota and Microsporidia. Current Biology *23*, 1548–1553.

Lill, R., and Kispal, G. (2000). Maturation of cellular Fe–S proteins: an essential function of mitochondria. Trends in Biochemical Sciences *25*, 352–356.

Netz, D.J.A., Stümpfig, M., Doré, C., Mühlenhoff, U., Pierik, A.J., and Lill, R. (2010). Tah18 transfers electrons to Dre2 in cytosolic iron-sulfur protein biogenesis. Nat Chem Biol *6*, 758–765.

Soler, N., Delagoutte, E., Miron, S., Facca, C., Baïlle, D., d'Autreaux, B., Craescu, G., Frapart, Y.-M., Mansuy, D., Baldacci, G., et al. (2011). Interaction between the reductase Tah18 and highly conserved Fe-S containing Dre2 C-terminus is essential for yeast viability. Mol Microbiol *82*, 54–67.