Functional consequences of reductive protein evolution in a minimal eukaryotic genome

Jason Jiang[1*], Rui Qu[1*], Maria Grigorescu[1], Winnie Zhao[1], and Aaron W. Reinke[1#]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
*These authors contributed equally
#Corresponding author
aaron.reinke@utoronto.ca

Abstract

Microsporidia are parasites with the smallest known eukaryotic genomes. The extent of protein loss in these organisms has been well documented, but much less is known about how compaction of microsporidia proteins affects their function. Taking a comparative genomic approach, we identified microsporidia orthologs of budding yeast proteins and show that these orthologs are enriched for essential yeast genes. We show that the median microsporidia protein is 21% shorter than its yeast counterpart and although extensive protein loss occurred after the divergence of microsporidia, reduced protein sizes were already present in microsporidian relatives. Microsporidia proteins are shorter through reduced domain lengths, diminished linker lengths, and domain loss, with 21% of microsporidia orthologs having lost domains present in yeast. On average, 34% of microsporidia orthologs have lost C-terminal residues essential for function in yeast, including 13 essential domains lost per genome. We also found that microsporidia display distinct phylogenetic patterns of domain loss, with losses occurring in a clade-specific manner. To investigate conservation of function, we used yeast complementation assays to test orthologs from several microsporidia species and their relative *Rozella allomycis*. These experiments reveal that most microsporidia proteins cannot complement their yeast orthologs, the ability to complement is about three-fold less than observed for *R. allomycis* orthologs, and proteins that do not complement are more reduced in length than their yeast orthologs. Altogether, our results demonstrate the drastic reduction of microsporidia proteins and show that these reductions have resulted in functional divergence from their fungal ancestors.

Introduction

The change from a free-living to a host-associated lifestyle is linked with genomic reduction and loss of protein-coding genes in bacteria, archaea, protists, and animals (Moran 2002; Lind et al. 2018; Mathur et al. 2019; Slyusarev et al. 2020). Microsporidia are an extreme example of genomic minimalization. These eukaryotic intracellular parasites are related to fungi and infect a wide range of animals and protists (Keeling 2009; Murareanu et al. 2021; Bojko et al. 2022). The microsporidia *Encephalitozoon cuniculi* was the first eukaryotic parasite sequenced, revealing a 2.9 megabase genome that codes for only ~2000 proteins, compared to ~6000

proteins in budding yeast (Katinka et al. 2001; Giaever and Nislow 2014). Microsporidia are highly reliant on their hosts and have undergone extensive gene loss, including in many metabolic genes (Keeling and Fast 2002; Nakjang et al. 2013; Wadi and Reinke 2020; Wadi et al. 2023). In addition to protein loss, the proteins themselves are on average ~15% shorter than their *Saccharomyces cerevisiae* orthologs (Katinka et al. 2001; Desjardins et al. 2015). Ribosomal structures from microsporidia have revealed the smallest eukaryotic ribosomes which show extreme levels of protein compaction, often with losses occurring at the N- or C-terminus (Barandun et al. 2019; Ehrenbolger et al. 2020; Nicholson et al. 2022). The proteosome has also been shown to have lost regulatory proteins and to have undergone compaction (Jespersen et al. 2022). Although the functional consequences of microsporidia protein compaction are mostly not known, it was speculated that these shorter proteins would result in reduced protein-protein interactions (Katinka et al. 2001). One example of protein reduction is the microsporidian leucine aminoacyl-tRNA synthetase which lacks a functional proofreading domain and protein synthesis at leucine residues have a high error rate (Melnikov et al. 2018).

The length and structure of proteins are constrained by evolutionary pressures. Proteins consist of independently folded domains which are the functional units of proteins, flanked by regions of amino acids known as linkers. Conservation of domain architecture has been observed across eukaryotic orthologs (Forslund et al. 2011). Proteins are shorter in prokaryotes than eukaryotes, largely explained by longer linker regions in eukaryotes as the domain sizes between the two kingdoms is similar (Brocchieri and Karlin 2005; Wang et al. 2005; Wang et al. 2011). Conserved proteins tend to be longer and longer proteins have more protein-protein interactions, suggesting a link between connectivity and protein length (Lipman et al. 2002; Warringer and Blomberg 2006). In eukaryotes, there is a slight tendency for smaller genomes to have larger average protein sizes, though microsporidia are an exception (Tiessen et al. 2012). There are fewer conserved protein domains in microsporidia than there are with other fungal species (Barrera et al. 2014). There are some examples of domain loss in microsporidia orthologs, but a comprehensive and systematic analysis of the changes that result in microsporidia protein compaction has not yet been performed (Heinz et al. 2012; Melnikov et al. 2018).

A powerful approach to explore if protein function is conserved between orthologs is to experimentally test if a diverged ortholog can replace the function of its counterpart. Studies using this approach have been facilitated by complete heterozygous knockout (hetKO) and temperature-sensitive collections in the ~1000 essential genes in the yeast *S. cerevisiae* (Giaever et al. 2002; Li et al. 2011; Giaever and Nislow 2014; Kofoed et al. 2015). Using these collections, the ability of human genes to replace deletions in essential orthologs has been systematically measured. About half of the ~400 human orthologs tested could rescue growth in their corresponding mutants. This result was only partly dependant on gene sequence identity and genes that function in the same process tended to display similar levels of replaceability (Kachroo et al. 2015). These types of experiments have demonstrated that conservation of protein function often extends across kingdoms, with 61% of *Escherichia coli* orthologs being able to complement yeast mutants (Kachroo et al. 2017). The ability of

orthologs from other species of yeast to complement *S. cerevisiae* deletions ranges from ~60-90% (Lai et al. 2023).

Yeast complementation has also been used to test the ability of microsporidia proteins to complement their yeast orthologs. Proteins from several species of microsporidia, including *E. cuniculi*, *Antonospora locustae*, and *Trachipleistophora hominis*, have been shown to rescue mutants in transcription, translation, the secretory pathway, and iron-sulphur cluster assembly (Hausmann et al. 2002; Hausmann et al. 2004; Slamovits et al. 2006; Upadhya et al. 2006; Goldberg et al. 2008; Freibert et al. 2017). There are also several examples of microsporidia proteins being unable to complement processes such as iron-sulphur cluster assembly or mitochondrial import (Goldberg et al. 2008; Waller et al. 2009; Freibert et al. 2017). As microsporidia genomes contain some of the smallest eukaryotic proteins, there has been interest in using complementation to understand how these minimal eukaryotic proteins function, but this approach has so far not been carried out systematically (Texier et al. 2013).

Recent ecological and sequencing efforts have identified relatives of the microsporidia (Wadi and Reinke 2020; Corsaro 2022). The closest known sister species to microsporidia is *Rozella allomycis,* which is a member of the Cryptomycota (James et al. 2013). The next closely related group of species are the early diverging microsporidia (also known as short-branched microsporidia) which are evolutionarily between *R. allomycis* and the canonical microsporidia (Bass et al. 2018). There are also the Metchnikovellids which are hyperparasites of gregarines and are basal to the canonical microsporidia (Galindo et al. 2018). Together these different microsporidian relatives provide an opportunity to understand the evolution of protein compaction in these highly reduced parasites.

A conserved microsporidia protein may be under selection to maintain the same functional ability as its budding yeast counterpart. Alternatively, the conserved microsporidia protein may be performing a different role or the context that the protein performs in has changed and thus the functional constraints the protein evolves under are different than in yeast. To distinguish between these possibilities, we used both bioinformatic and experimental approaches to study microsporidia protein evolution. We identified single-copy orthologs between microsporidia species and yeast. This analysis revealed a set of mostly essential orthologs with diminished proteins sizes and many of the lost residues are functional in their yeast orthologs. Moreover, domain loss was highly prevalent, occurring in 21% of microsporidia orthologs, with many of these lost domains essential in yeast. Additionally, we observed that domain losses occurred in a clade-specific manners in microsporidia. Using yeast complementation assays, we directly tested the function of microsporidia orthologs, observing that microsporidia proteins are significantly less likely to complement than their sister group *R. allomycis*. Our study suggests that microsporidia proteins have undergone extensive loss of function compared to their yeast orthologs and provides a system for further understanding protein evolution in a minimal eukaryotic genome.

Results

Microsporidia share a predominantly essential and reduced in length set of single-copy orthologs with *S. cerevisiae*

To compare the evolutionary properties of microsporidia proteins, we first identified orthologs shared between microsporidia species and the yeast *S. cerevisiae* using OrthoFinder (Emms and Kelly 2019). We analyzed 41 microsporidia species which we refer to as the canonical microsporidia, two species of the Metchnikovellids (Mikhailov et al. 2017; Bass et al. 2018; Galindo et al. 2018; Wadi and Reinke 2020), the early diverging microsporidia species *Paramicrosporidium saccamoebae* and *Mitosporidium daphniae* (James et al. 2013; Haag et al. 2014; Quandt et al. 2017), and the Cryptomycota species *R. allomycis*. For comparison, we also included six eukaryotic outgroup species (Supplementary table S1). We only considered single-copy orthologs as paralogous gene relationships display divergent function (Forslund et al. 2011; Soria et al. 2014) (fig. 1A and methods). On average, microsporidia species shared 438 single-copy orthologs with yeast (Supplementary fig. S1 and Supplementary table S2). The majority (61%) of these conserved microsporidia proteins are orthologous to essential yeast genes. In comparison, *R. allomycis* contained over 1000 orthologous genes with ~40% of them being essential in yeast (Supplementary fig. S1 and Supplementary table S2).

Previous reports on microsporidia genomes have reported that microsporidia proteins are shorter than their yeast orthologs (Katinka et al. 2001; Desjardins et al. 2015). To see how widely conserved this phenomenon is, we compared the length of each microsporidian ortholog to its yeast counterpart. We find that the median canonical microsporidia ortholog is 21% shorter than its yeast counterpart, compared to the median eukaryotic outgroup protein being less than 2% shorter (fig. 2A and Supplementary table S2). In both microsporidia and the outgroup, orthologs of non-essential proteins are slightly shorter than those of essential proteins (fig. 2C). Most microsporidia proteins (73%) are less than 90% the length of their yeast counterparts. In contrast, only a quarter of the outgroup orthologs are this reduced (Supplementary fig. 2). *R. allomycis* and the early diverging microsporidia represent an intermediate, as slightly over half their orthologs are less than 90% the length of their yeast counterparts (fig. 2A and Supplementary fig. 2). Microsporidia orthologs are also diverged in sequence content from their yeast orthologs, sharing only 36% sequence identity on average. The other groups are also diverged, between 38-40% sequence identity. For all groups examined, essential orthologs have a slightly higher average sequence identity than non-essential orthologs (Supplementary fig. S3A and table S2). Although microsporidia orthologs are both reduced in length and diverged in sequence identity, there is poor correlation between these two attributes (Supplementary fig. S3B).

Conserved microsporidia proteins have undergone widespread domain loss

To better understand how protein reduction affected the structure of microsporidia proteins, we compared domain architectures between microsporidia and yeast single-copy orthologs. Three major domain architectural

141 changes were identified: domain loss, domain gain, and domain swap (fig. 1B and supplementary table S3).
142 Domain architecture between most ortholog pairs was conserved, with 72% of microsporidia orthologs having
143 identical domain architectures to their yeast orthologs (fig. 2B). *R. allomycis* and early-diverging microsporidia
144 orthologs were conserved to a similar extent (69%) and the eukaryotic outgroup orthologs were the most
145 conserved (82%). In canonical microsporidia, domain loss was the most prevalent architectural change,
146 occurring in 21% of orthologs, which is about twice as much in *R. allomycis* and the early diverging species.
147 Outgroup orthologs showed a domain loss rate of only 2.5% (supplementary table S3). We also observed higher
148 rates of gained domains in *R. allomycis*, early diverging microsporidians, and the outgroup compared to either
149 the canonical microsporidia or the metchnikovellids. Domain architecture changes occurred to similar extents in
150 both essential and non-essential yeast orthologs (fig. 2B).
151

152 In addition to domain loss, microsporidia orthologs could become shorter through length reduction of their
153 domains, linker sequences, or both. To ascertain which of these possibilities has occurred, we examined
154 microsporidia orthologs with a conserved domain structure. Microsporidia domains have a median of 89% of
155 their corresponding yeast ortholog domain lengths. An even greater decrease was seen in linker lengths with
156 the median microsporidian linkers being 70% of their corresponding yeast ortholog linker lengths. In both cases,
157 only small differences were observed between essential and non-essential orthologs. Outgroup ortholog domain
158 and linker lengths were less reduced than in microsporidia, *R. allomycis,* and early-diverging microsporidia.
159 Diverging species were intermediary (fig. 2C and Supplementary table S4).

160 **Shortened microsporidia orthologs have lost residues essential for function in yeast**

161

162 Although microsporidia proteins have been observed to be shorter than their yeast orthologs, the functional
163 consequences of these losses are not known. One possibility is that microsporidia proteins are just losing
164 residues that are non-essential in yeast. To determine if microsporidia orthologs have lost residues that are
165 essential in their yeast counterparts, we analyzed data from a previous study that used CRISPR-Cas9 to insert
166 premature stop codons (PTCs) into essential yeast genes (Sadhu et al., 2018). We calculated whether the
167 residues that were lost at the C-terminus in microsporidia orthologs were either essential, non-essential, or
168 ambiguous in yeast (fig. 3A and Supplementary table S5). We observed that the number of essential residues
169 lost in microsporidia orthologs was higher than in the outgroup species (fig. 3B). Of the proteins that are shorter
170 than their yeast orthologs, 37% have lost essential residues compared to 22% of outgroup proteins (fig. 3C). An
171 even more striking difference is observed if all proteins are considered; 34% of all microsporidia have lost
172 essential residues compared to 12% of outgroup proteins. In addition to losing essential residues, the median
173 microsporidia ortholog has lost all dispensable residues, with only ~70% of outgroup orthologs having done so
174 (fig. 3D). Together these results suggest that residues necessary for function in yeast have been lost from the
175 C-terminus of microsporidia orthologs.

176 Protein domains in yeast that are lost in their microsporidia orthologs might be carrying out an essential function.
177 Using the premature stop codon data, we classified each C-terminal domain as essential, dispensable, or
178 ambiguous (fig. 4A and Supplementary table S6) (Sadhu et al. 2018). We observed that on average each species
179 of microsporidia has lost 13 domains that are essential in yeast, compared to 1 domain per eukaryotic outgroup
180 species (fig. 4B).
181
182 **Domain architectural changes in microsporidia orthologs of yeast proteins occurred early in evolution**
183 **in clade-specific patterns**
184
185 We then sought to determine whether domain architectural changes in microsporidia orthologs followed any
186 phylogenetic patterns. First, we clustered orthologs by their domain architectural classifications of being
187 conserved, lost domains, or gained/swapped domains (fig. 5A). Clustering revealed a core set of yeast genes
188 conserved across microsporidia that have mostly conserved domain architectures. In addition, some orthologs
189 lost in canonical microsporidia were still present in microsporidian ancestors with conserved domains. Moreover,
190 there was a prominent cluster of genes that lost domains in most microsporidia species. However, there were
191 also incidences of individual species and clade-restricted domain losses and gains.
192
193 As groups of genes seemed to lose domains based on their phylogenetic relationships, we investigated if there
194 were any phylogenetic patterns to the lost domains themselves. We clustered microsporidia species using lost
195 domains in their yeast orthologs. Indeed, microsporidia have undergone patterns of domain loss consistent with
196 their phylogenetic groupings (fig. 5B and supplementary table S7). For instance, species were separated into
197 their broader evolutionary clades, with canonical microsporidia and microsporidia relatives grouping into distinct
198 clusters. Moreover, we observed genera-specific patterns of domain loss, such as with *Nematocida* and
199 *Encephalitozoon* forming their own distinct clusters.
200
201 **Most microsporidia orthologs do not complement essential yeast deletions**
202
203 To experimentally determine the ability of microsporidia proteins to function in place of their *S. cerevisiae*
204 orthologs, we used yeast complementation assays. To test if microsporidia orthologs rescue growth of essential
205 yeast genes, we used diploid strains of yeast that contained a heterologous deletion. We transformed these
206 strains with plasmids containing constitutive expression of each cloned ortholog that corresponded to the
207 deletion. We then sporulated the yeast strains and performed random spore analysis by plating the resulting
208 haploid cells on selective media to determine if the microsporidia gene could rescue the growth of the haploid
209 deletion strain (fig. 6A). We chose representative genes and tested orthologs from three representative
210 microsporidia species which belong to different clades (Bojko et al. 2022); *N. parisii*, *T. hominis*, and *E. cuniculi*.
211 We also included the closest relative of microsporidia, *R. allomycis*. In total we tested 29 yeast deletions for a

212 total of 82 tests (Supplementary table S8). We found that *R. allomycis* had the highest complementation rate,
213 with 8/27 genes successfully complementing (30%). In contrast, complementation by microsporidia genes
214 ranged from 8-11% and are significantly less likely to complement than *R. allomycis* (fig. 6B). Even for genes
215 that *R. allomycis* could complement, the microsporidian complementation rate was only 33% and microsporidia
216 complementation was only ever observed if the *R. allomycis* ortholog could also complement (Supplementary
217 table S9).
218
219 We then compared length and domain conservation in the proteins that were functional for complementation. *R.*
220 *allomycis* proteins that complemented have an average length 96% that of their yeast orthologs. In contrast, *R.*
221 *allomycis* proteins that could not complement were significantly shorter with an average length 86% that of their
222 yeast orthologs (Fig. 6C). Microsporidia proteins that complemented had an average length 94% that of their
223 yeast orthologs, while those that could not complement had an average length 76% that of their yeast orthologs
224 (Fig. 6C and Supplementary table S9). The most reduced *R. allomycis* protein that complemented was 85% the
225 length of its yeast ortholog and the most reduced microsporidia protein that complemented was 88% the length
226 of its yeast ortholog.
227
228 One of the reduced microsporidia proteins that complemented was the ribosomal stalk protein RPP0 for which
229 we observed complementation for the *R. allomycis*, *E. cuniculi*, and *N. parisii* orthologs (Supplementary tables
230 S8 and S9). The C-terminal part of the yeast RPP0 protein contains regions which are necessary for facilitating
231 interactions with other ribosomal proteins and C-terminal truncations of RPP0 diminish yeast growth
232 (Supplementary fig. S4A) (Santos and Ballesta 1995). These regions are missing in microsporidia orthologs. To
233 determine if yeast RPP0 that was missing C-terminal residues could rescue to a similar extent, we generated
234 RPP0 truncation mutants and observed that the two microsporidia, the full-length yeast, and the truncated yeast
235 versions could all support growth by the same amount (Supplementary fig. S4B).
236
237 We then determined the reported rate of complementation of yeast deletions with microsporidia proteins by
238 compiling a list of previously reported complementation assays (Supplementary table S10). The rate reported in
239 published articles is 9 out of 25 (36%) proteins complementing, and in theses 1 out of 7 (14%), for an overall
240 reported complementation rate of 31%. However, some of these proteins that complemented were cloned using
241 exogenous mitochondrial targeting signals, and if these cases are omitted, 6/28 (21%) complement
242 (Supplementary fig. S5A). We then compared the length of the microsporidia proteins tested to their yeast
243 orthologs. Proteins that don't complement are more reduced (Supplementary fig. S5B). The combined
244 complementation data of our study and previous reported results shows that microsporidia proteins that
245 complement are significantly less reduced (Supplementary fig. S5C). Altogether, our analysis indicates that most
246 microsporidia proteins cannot complement deletions of their yeast orthologs and that more reduced proteins are
247 less likely to complement.

Discussion

Reductive evolution is a hallmark of intracellular microbes (Moran 2002). Using both computational and experimental methods we show that microsporidia proteins have undergone functional losses compared to their budding yeast counterparts. We show that microsporidia proteins are unique to other eukaryotic proteins in several ways. Microsporidia orthologs are enriched for essential genes, have undergone extreme reduction in length, and have lost residues and domains essential in yeast. The loss of functional residues and domains we observe is an underestimation, as only the impact of C-terminal residues has been determined, and it is likely that many of the residues we classify as ambiguous are necessary for function. Together our results suggest that microsporidia proteins have experienced widespread loss of the corresponding residues necessary for function of yeast proteins. One limitation of our approach is orthologs are identified based on sequence identity, which can fail to identify highly divergent microsporidia orthologs (Nakjang et al. 2013). Approaches in the future using structural based detection may be able to overcome this limitation (Mascarenhas dos Santos et al. 2022; Svedberg et al. 2023).

We observe that most microsporidia proteins cannot complement essential yeast deletions and that microsporidia proteins complement at a lower rate than has been observed for other species. Yeast complementation assays have been used on orthologs from a variety of different species with complementation rates ranging from ~50-90% (Kachroo et al. 2015; Kachroo et al. 2017; Lai et al. 2023). Although we only tested a subset of microsporidia genes for their ability to complement yeast deletions, we provide ~3 times more complementation tests than have been previously reported for microsporidia. The overall rate of complementation we see is somewhat lower than what has been reported in the literature, though our study includes more highly reduced proteins and is likely more representative of the ability of microsporidia proteins to rescue yeast deletions. One reason that microsporidia proteins display low complementation is not localizing correctly in yeast and adding a tag has been shown to correct localization and rescue function (Goldberg et al. 2008). Another possibility is that expression levels are too low, but many reported microsporidia complementation tests used multi-copy plasmids and strong reporters and yet didn't observe complementation (Supplementary table S10). Another explanation for failure to complement is that an essential interaction is no longer maintained between the microsporidia ortholog and a yeast protein. It could be that microsporidia proteins have rapidly co-evolved and the microsporidia protein can't function in the yeast context due to differences in their interacting partners. This type of evolution was recently shown for the anaphase-promoting complex, where the replacement of multiple genes from other yeast species at the same time was necessary to restore function (Lai et al. 2023). Alternatively, the microsporidia protein could have evolved to no longer carry out the essential function of its yeast ortholog. Further investigation will be necessary to understand the explanation for each orthologs' inability to complement.

284

285 Our results suggest that microsporidia evolution has resulted in progressive reduction and loss of function
286 compared to other eukaryotes. A loss of both conserved genes and metabolic pathways has been observed in
287 *R. allomycis* compared to other fungi, with more losses occurring in the early diverging microsporidia species
288 (James et al. 2013; Haag et al. 2014; Quandt et al. 2017; Wadi et al. 2023). Additional losses were reported in
289 the metchnikovellids and canonical microsporidia, with gene loss occurring to similar extents in these groups
290 (Mikhailov et al. 2017; Galindo et al. 2018; Wadi et al. 2023). Consistent with what has been shown for
291 conservation of proteins, we observe that *R. allomycis* and the early diverging microsporidia are more reduced
292 than other eukaryotes, but not to the same extent as the metchnikovellids and the canonical microsporidia. We
293 observe similar trends for domain loss and for residues that are functional in yeast. Together, our data suggests
294 that protein compaction and domain loss had already occurred before the divergence of the canonical
295 microsporidia. In complementation assays, we observe that *R. allomycis* can rescue the function of deletions at
296 a rate that is higher than microsporidia, but less than species of yeast, providing further evidence that although
297 *R. allomycis* is a parasitic species, it represents an evolutionary intermediate between free-living fungi and
298 microsporidia.

299

300 Although microsporidia were discovered over 160 years ago, work on defining their core functional mechanisms
301 has largely been unexplored. Genomic approaches have been extremely useful at defining what genes are
302 present, but there is a lack of genetic tools available to test their function (Katinka et al. 2001; Reinke and Troemel
303 2015; Wadi and Reinke 2020). Recent high resolution structural approaches have documented how
304 microsporidia proteins and complexes undergo compaction (Barandun et al. 2019; Ehrenbolger et al. 2020;
305 Jespersen et al. 2022; Nicholson et al. 2022). Biochemical and ancestral reconstruction approaches have also
306 been extremely valuable to understanding functional evolution of proteins (Goldberg et al. 2008; Freibert et al.
307 2017; Dean et al. 2018). The use of yeast as a heterologous system to study these proteins provides a powerful
308 approach to systemically determining functional conservation (Kachroo et al. 2015). Our finding that some *R.*
309 *allomycis* proteins can complement yeast deletions, but not the corresponding microsporidia orthologs, provides
310 a comparative system to further explore evolution of microsporidia protein function. Using approaches such as
311 chimeras and determining differences such as localization and protein-protein interactions will be useful for
312 exploring how microsporidia orthologs have functionally diverged.

313

314 **Materials and Methods**

315 **Identification of *Saccharomyces cerevisiae* single-copy orthologs and alignment of domain**
316 **architectures**

317 Orthogroups between proteomes were identified with OrthoFinder v2.5.4 (Emms and Kelly 2019), using BLAST
318 2.5.0+ (Johnson et al. 2008) for the all-versus-all search step. A total of 53 proteomes were used, including
319 microsporidia genomes, microsporidia relatives, *S. cerevisiae*, and outgroups. The proteomes used and their

accession numbers are listed in Supplementary table S1. After the initial OrthoFinder-step, orthologs from 8 species were excluded for being judged to be low quality assemblies (Wadi et al. 2023). Orthogroups containing a single-copy *S. cerevisiae* ortholog and at least one single-copy ortholog from any other species were identified for downstream analyses. *S. cerevisiae* orthologs were classified as essential or non-essential using a previously published list of essential yeast genes (Kofoed et al. 2015).

Protein domains in each ortholog were assigned using hmmscan from HMMER 3.3.2 (Eddy 2011) and Pfam 35.0 (Mistry et al. 2021). Pfam family-specific bit-score cut-offs for protein domain families were used over fixed E-value thresholds, due to previously demonstrated improvement benefits in using these bit-score cutoffs over E-values (Punta et al. 2012). Overlapping domain assignments from hmmscan were resolved with cath-resolve-hits 0.16.10 (Lewis et al. 2019) to obtain the final domain architectures for each ortholog. Within each orthogroup, the *S. cerevisiae* single-copy ortholog was aligned with all other single-copy ortholog in the orthogroup. To facilitate alignment, all domains in domain architectures were first mapped back to their Pfam clans, as domain families within clans have conserved function (Forslund et al. 2011; Mistry et al. 2021). Within each domain architecture, the clans were converted to unique letter representations. Finally, the letter representations of clans within each ortholog domain architecture were aligned with the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), as has been previously done for domain architectural alignments (Forslund et al. 2011), using a match score of 0, a mismatch penalty of -3, and a gap penalty of -10.

Identification of domain architectural changes from domain architecture alignments

We defined three categories of domain architectural changes in orthologs compared to their *S. cerevisiae* orthologs: domain losses, domain gains, and domain swaps. Domain losses in species orthologs were inferred as a gap in the alignment of the species domain architecture compared to its *S. cerevisiae* ortholog. Domain gains were inferred as gaps in the aligned *S. cerevisiae* ortholog domain architecture compared to its species ortholog. Domain swaps were inferred as mismatching positions between the aligned domain architectures. For species-*S. cerevisiae* ortholog pairs where the species ortholog lacked domain assignments, we considered the species ortholog to have lost all the domains present in its *S. cerevisiae* ortholog. Conversely, if the *S. cerevisiae* ortholog lacked domain assignments, we considered the species ortholog to have gained all its domains relative to the *S. cerevisiae* ortholog. To account for erroneous missing domain classifications, we excluded *S. cerevisiae* -species ortholog pairs with domain loss where the species ortholog was not shorter than its *S. cerevisiae* ortholog by at least 85% of the length of all of its lost domains from *S. cerevisiae*. We employed this heuristic filter to exclude instances of domain loss classified by our pipeline that were due to high sequence divergence in the species ortholog sequence from *S. cerevisiae*. Such *S. cerevisiae* -species ortholog pairs were excluded from all downstream analyses.

Comparison of domain and linker lengths between microsporidia and yeast single-copy orthologs

353 Total domain length for an ortholog was calculated as the summed lengths of all domains within the ortholog.
354 For orthologs without domain assignments, the total domain length was defined as zero. We excluded ortholog
355 pairs from orthogroups where neither the other species nor *S. cerevisiae* ortholog had domain assignments.
356 Linker lengths were defined as the total protein length for an ortholog with total domain length subtracted. To
357 compare total domain and linker lengths between ortholog pairs, we divided the length in the other species by
358 the length in the *S. cerevisiae* ortholog.

359 **Identification and classification of lost C-terminal amino acids in single-copy yeast-essential genes**

360 First, we classified each amino acid within all essential *S. cerevisiae* genes as either essential, dispensable or
361 ambiguous for *S. cerevisiae* viability, using a dataset of premature stop codons (PTCs) inserted into all essential
362 yeast genes and their effects on yeast viability (Sadhu et al. 2018). All amino acids coming at or before a yeast-
363 lethal PTC were classified as essential. All amino acids coming at or after a tolerated PTC were classified as
364 dispensable. However, amino acids were classified as ambiguous in the following cases: (1) amino acids coming
365 after the C-terminus PTC in a gene and the C-terminus PTC is lethal, (2) amino acids coming before the N-
366 terminus PTC and the N-terminus PTC is tolerated, and (3) amino acids between a lethal and tolerated PTC.
367 Amino acids in (1) were excluded, as it is unclear whether amino acids after the C-terminus lethal PTC are
368 essential or dispensable. Amino acids in (2) were excluded as it is unclear if essential amino acids exist upstream
369 of the N-terminus tolerated PTC. Amino acids in (3) were excluded as it is unclear if amino acids exist upstream
370 of the tolerated PTC, in the region between the lethal and tolerated PTC.

371 To identify lost C-terminal amino acids in orthologs to yeast, we first identified yeast-essential orthologs that were
372 shorter than their yeast equivalents. Such ortholog pairs were aligned with the Smith-Waterman local alignment
373 algorithm, using the BLOSUM45 substitution matrix for microsporidia orthologs to account for their high
374 sequence divergence (Nakjang et al. 2013), and the BLOSUM62 matrix for all other species. Lost C-terminal
375 residues were defined as residues in the yeast ortholog coming after the end of the local alignment and were
376 classified as described above.

377 **Classification of dispensability of lost domains in single-copy orthologs to yeast**

378 We classified domains in all essential *S. cerevisiae* genes as “essential”, “dispensable” or “ambiguous” using the
379 PTC dataset in essential yeast genes and their effects on yeast viability (Sadhu et al. 2018). Within each essential
380 yeast gene, we indicated the C-terminal most PTC as the “critical PTC”. Essential domains were classified as
381 domains for which the critical PTC appears upstream of the domain end, suggesting that truncation/loss of the
382 domain by the critical PTC led to yeast mortality. Moreover, we also classified domains with the critical PTC
383 immediately downstream of its end as essential, to account for cases in which the real boundaries of the domain
384 extend beyond the predicted boundaries and the domain was actually disrupted by the critical PTC. Dispensable
385 domains were classified as domains with a dispensable PTC upstream of its start, suggesting that the domain
386 could be completely removed without affecting yeast fitness. All other domains were classified as “ambiguous”.

387 All lost domains in *S. cerevisiae* -essential orthologs were classified, per the classification of the domain in their
388 corresponding *S. cerevisiae* ortholog.

389 **Hierarchical clustering of single-copy orthologs to yeast by domain architecture**

390 First, a table was constructed of all single-copy yeast orthologs with at least one single-copy ortholog from
391 another species. Columns of the table were yeast orthologs, rows were the non-yeast species (microsporidia
392 and outgroups) and cells held information on the domain architecture conservation between the yeast ortholog
393 and the single-copy ortholog from the species in that row. Domain architecture conservation between yeast-
394 species ortholog pairs were classified as follows. Ortholog pairs with identical domain architectures were
395 classified as "Conserved domain architecture". Ortholog pairs where the species ortholog had lost one or more
396 domain were classified as "Lost domain(s)". Ortholog pairs having gained or swapped one or more domains
397 were classified as "Gained/swapped domain(s)". If a species ortholog had both lost and swapped domains, we
398 classified it as "Lost domain(s)". In cases where a species did not have a single-copy ortholog to a particular
399 yeast ortholog, we classified that cell as "Ortholog not conserved". With this table, a distance matrix was
400 calculated for the *S. cerevisiae* orthologs, based on their dissimilarity in domain architecture conservation across
401 species. Gower's distance was used to calculate distances for these categorical variables, using the daisy
402 function from the cluster v2.1.2 R package (Maechler et al. 2022). Yeast orthologs were then clustered by
403 similarities in domain architecture conservation patterns across species, using the agnes function from the
404 cluster R package for agglomerative hierarchical clustering with the distance matrix.

405 **Generation of t-SNE plot of ortholog lost domains.**

406 First, a list of all unique lost Pfam clans was collected for each species, using the aligned ortholog domain
407 architectures for each species and their yeast orthologs. A dissimilarity matrix was constructed between each
408 species and their set of unique clan losses using Gower's distance, with the daisy function from R package
409 cluster v2.1.2 (Maechler et al. 2022). Using this matrix, species were then clustered by their similarity in lost
410 clans using k-medoids clustering, using the pam function from cluster. The optimal number of clusters (k) for k-
411 medoids clustering was selected by clustering with k = 2 to k = number of species - 1, and selecting the k-value
412 that maximizes silhouette width of the clustering (Rousseeuw 1987), which is a measure of how well items fit in
413 their clusters. The clustered species were visualized using t-SNE, passing the dissimilarity matrix into the Rtsne
414 function from the Rtsne v0.15 R package (Krijthe 2015). Maximum iterations for t-SNE was set to a high value
415 of 5000, to ensure convergence to stable t-SNE plots. Ideal perplexity was chosen by iterating from 5 to (number
416 of species – 1 / 3), and manually inspecting the resulting t-SNE graphs. This range of perplexities was chosen
417 based on the recommended range of perplexities of 5 – 50 (Maaten and Hinton 2008), and from the
418 documentation of Rtsne stating a maximum perplexity of (number of rows in data – 1 / 3).

419 **Cloning of microsporidia genes**

420

421 Genes to be cloned were chosen using two strategies. For the first strategy we chose 14 genes where there
422 were *R. allomycis*, *T. hominis,* and *N. parisii* orthologs. For 7 of these genes the length ratio of the *N. parisii*
423 protein compared to its yeast ortholog had a ratio greater than 0.99. For the other 7 genes this ratio was less
424 than 0.9. For the second strategy, 15 genes chosen at random where there were *R. allomycis, T. hominis,* and
425 *E. cuniculi* orthologs. All genes chosen were required to encode for *R. allomycis* proteins that are less than 479
426 amino acids, to facilitate cloning using gene synthesis. Genes from *Nematocida parisii*, *Encephalitozoon cuniculi*,
427 and *Saccharomyces cerevisiae* were PCR amplified from *N. parisii* ERTm1*, E. cuniculi* E3, and S. cerevisiae
428 BY4741 genomic DNA, respectively. Using ApE (Davis and Jorgensen 2022), primers for PCR were designed
429 for each gene of interest, with a target melting temperature of ~55 °C. To facilitate Gateway Cloning, partial attB
430 sequences were appended to the ends of the primers. The 5' end of the forward primer contained the sequence
431 ACAAAAAAGCAGGCTCA, whereas the 3' end of the reverse primer contained the sequence
432 GGGGACCACTTTGTACAAGAAAGCTGGGTT. Primers were designed from the start codon of every gene until
433 the last codon before the stop. Next, a two-step PCR was conducted. In the first step, the gene of interest was
434 amplified from genomic DNA using the gene-specific primers and Phusion polymerase. The second step
435 amplified the product of the first step with attB primers, which appended the complete attB1 sequence to the 5'
436 end of the gene and attB2 sequence to the 3' end. Amplicons from the two-step PCR were PEG-purified and
437 successful amplification of genes was verified by agarose gel electrophoresis. Genes from *R. allomycis, T.*
438 *hominis,* and some *E. cuniculi* genes were ordered as codon optimized synthesized genes from IDT. These
439 genes were ordered with attB1 and attB2 flanking sequences.
440
441 The PCR products were inserted into the pDONR221 entry vector in the BP reaction of the Gateway cloning
442 system and transformed into competent DH5-α *E. coli* cells to obtain entry clones. The extracted entry clones
443 were then inserted into the pAG416GPD-+6Stop (Laurent et al. 2020) destination vector by Gateway LR reaction
444 and transformed into competent E. coli to obtain expression clones. All entry and expression clones were verified
445 by restriction digests and Sanger sequencing. All oligos and gene sequences are shown in Supplementary table
446 S11.
447
448 **Transformation of expression clones into mutant yeast strains**
449
450 Expression clones were transformed into yeast strains as described previously (Reece-Hoyes and Walhout
451 2018). The empty destination vector, pAG416-GPD-ccdB+6Stop, was also transformed into cells as a negative
452 control. Yeast cells from transformed hetKO cells were plated on -uracil dextrose + G418 plates at 30°C for 2 -
453 3 days, to select for both successful transformation and presence of the G418-resistance knockout cassette.
454 Yeast cells from transformed temperature-sensitive strains were plated onto -uracil dextrose plates and grown
455 at 25 °C for 5 days.
456

457 **Heterozygous knockout strain complementation testing**

458

459 The ability of yeast and microsporidian orthologs to complement their corresponding heterozygous-knockout
460 yeast strains was tested with random spore analysis (protocol adapted from:
461 https://openwetware.org/wiki/Springer_Lab:_Random_Spore_Analysis). Single colonies of transformed yeast
462 cells were suspended in 2 mL of 1% potassium acetate, and sporulated for 5 days in a rotary shaker at 30°C.
463 Successful sporulation was confirmed with the observation of tetrads using microscopy. Cells were washed and
464 resuspended in 500µL milli-Q water, 25µL 5mg/mL 20T Zymolyase, and 5µL β-mercaptoethanol. The spore
465 mixture was placed on a nutating mixer overnight at 30°C. Tetrads were disrupted the following day by the
466 addition of Zirconium beads and shaking on a Disruptor Genie at 3000 RPM for 10 minutes. Cells were then
467 isolated with centrifugation for 10 minutes at 10,000 RCF. To further remove any diploids, the Zymolyase
468 treatment was repeated with 20mg/mL 20T Zymolyase, followed by Zirconium bead disruption the next day. The
469 cells were then plated onto -uracil dextrose plates and incubated at 25°C for 5 days to select for colonies
470 containing the plasmid of interest. Colonies were replica plated onto YPD+G418 and 5-FOA plates to test for
471 complementation. Haploid yeast cells that complement grow on YPD+G418 but not 5-FOA, and are expected to
472 grow in a ~1:1 ratio with haploid yeast cells with the wild-type gene copy, which grow on 5-FOA but not
473 YPD+G418. Cells that could grow on both plates were considered diploids. The corresponding empty vector
474 negative control was also tested for any genes that showed complementation across all strains. Random spore
475 analysis was repeated once with each strain, to generate at least two biological replicates.

476

477 **Testing RPP0 protein functionality via dilution assays**

478

479 RPP0 truncation mutants were cloned as described above. The yeast truncation mutants, and the *N. parisii* and
480 *E. cuniculi* orthologs were transformed into the RPP0 hetKO strain and sporulated as described above. Dilution
481 assays were performed as described previously (Petropavlovskiy et al. 2020). For each plasmid, three
482 complementing haploid mutant colonies were picked and grown separately overnight in -uracil dextrose medium
483 at 25 °C, generating three different biological replicates. An empty vector negative control was picked from wild-
484 type haploid colonies. Cell concentrations of each culture were equalized to an OD600 of 1.0 and added to the
485 first row of a 96-well plate. Five 5X dilutions were then performed down the rows of the plate. Next, using a
486 spotting tool, equal amounts of each sample were spotted onto a -uracil dextrose plate. Plates were then
487 incubated for 2 days at 25 °C, 35 °C, or 37 °C. Images were taken of each plate after the 2-day incubation. Then,
488 using ImageJ (Schneider et al. 2012) the images were adjusted so that the white colonies showed sharp contrast
489 against the dark background. Then, the amount of growth of each sample and dilution was determined by
490 measuring the grey value of each. The grey value was also measured for five points amongst the background,
491 which was then averaged and subtracted from the grey value measurements for each of the samples. A mean
492 value of each condition was then generated by averaging all the grey value measurements of the condition from

each biological and technical replicate. Across each temperature, the mean values for all strains were then normalized against the value of the full-length yeast gene.

Data and code availability

Code for analyses was written in R (R Development Core Team 2019) and Bash. Scripts were organized into a pipeline using Snakemake (Köster and Rahmann 2012). All software needed to run the pipeline were deployed with a Singularity X.Y.Z container. 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.

Acknowledgements

We thank Qingyuan Huang, Edward James, Yin Chen Wan, and Jonathan Tersigni for providing helpful comments on the manuscript. We thank Brenda Andrews for the kind gift of yeast mutant strains. We thank Grant Brown for the kind gift of the pAG416-GPD-ccdB+6Stop vector. We thank Louis Weiss for providing *E. cuniculi* DNA. We thank Aashiq Kachroo for providing protocols and advice. We thank Alexander Ensminger, Marc Meneghini, and their labs for advice on working with budding yeast. This work was supported by a Canadian Institutes of Health Research grant no. 400784 (to A.R.), an Alfred P. Sloan Research Fellowship FG2019-12040 (to A.R.). and the Natural Sciences and Engineering Research Council of Canada Undergraduate Student Research Awards to J.J., and Undergraduate Summer Research Program awards to R.Q. and M.G.

Competing Interests: The authors declare that they have no competing interests.

Figures and legends

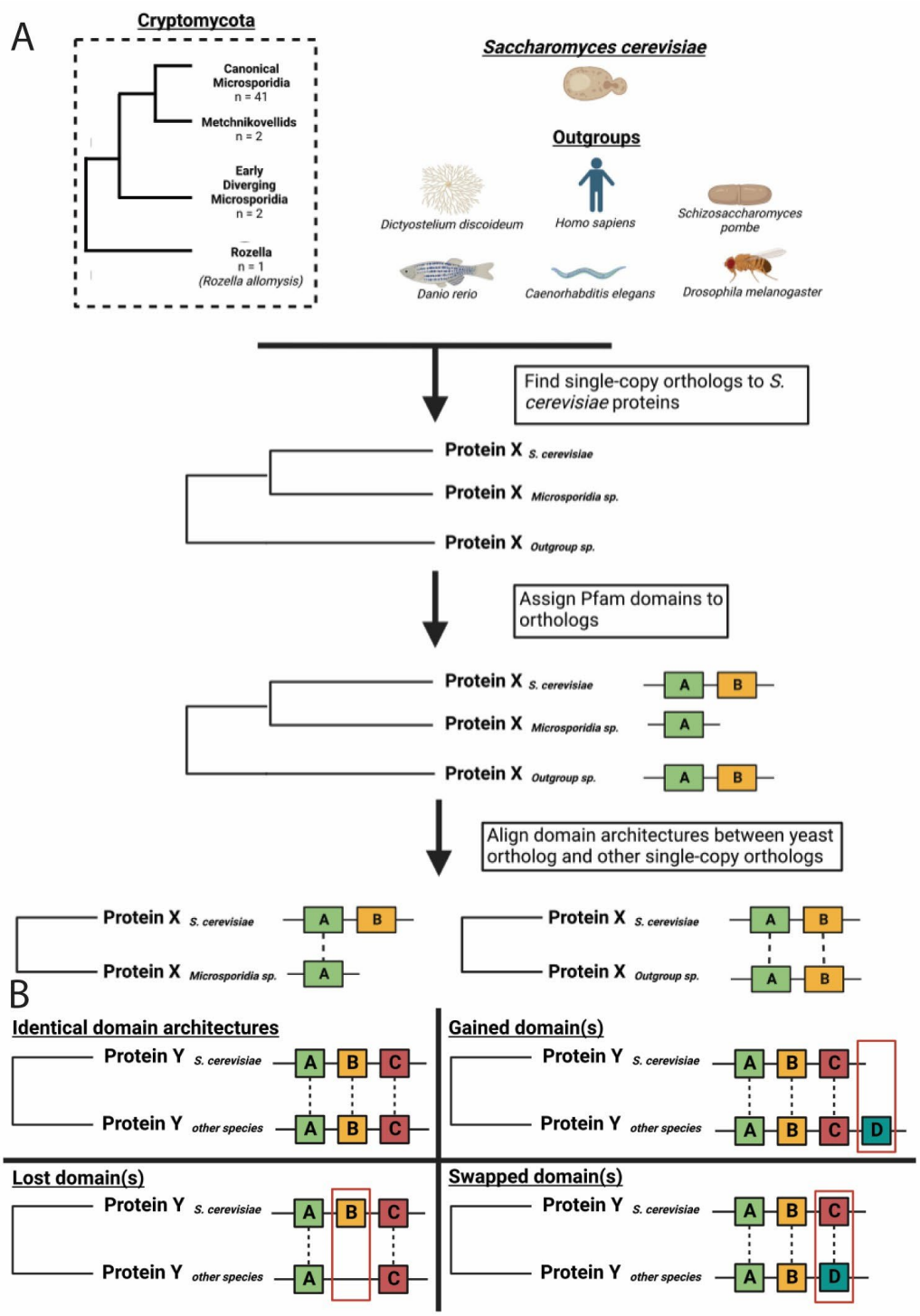


FIG. 1. Schematic overview of identifying single-copy orthologs and aligning orthologous domain architectures. (A) Computational workflow for identification of single-copy orthologs shared between *S. cerevisiae* and microsporidian or outgroup species. Shared single-copy orthologs are identified using Orthofinder and then Pfam domains are assigned. These domain architectures are then aligned. (B) Once domains are aligned, they are then compared to identify cases where domains are either conserved, gained in the other species compared to yeast, lost in the other species compared to yeast, or swapped, with a different domain in the other species compared to yeast.

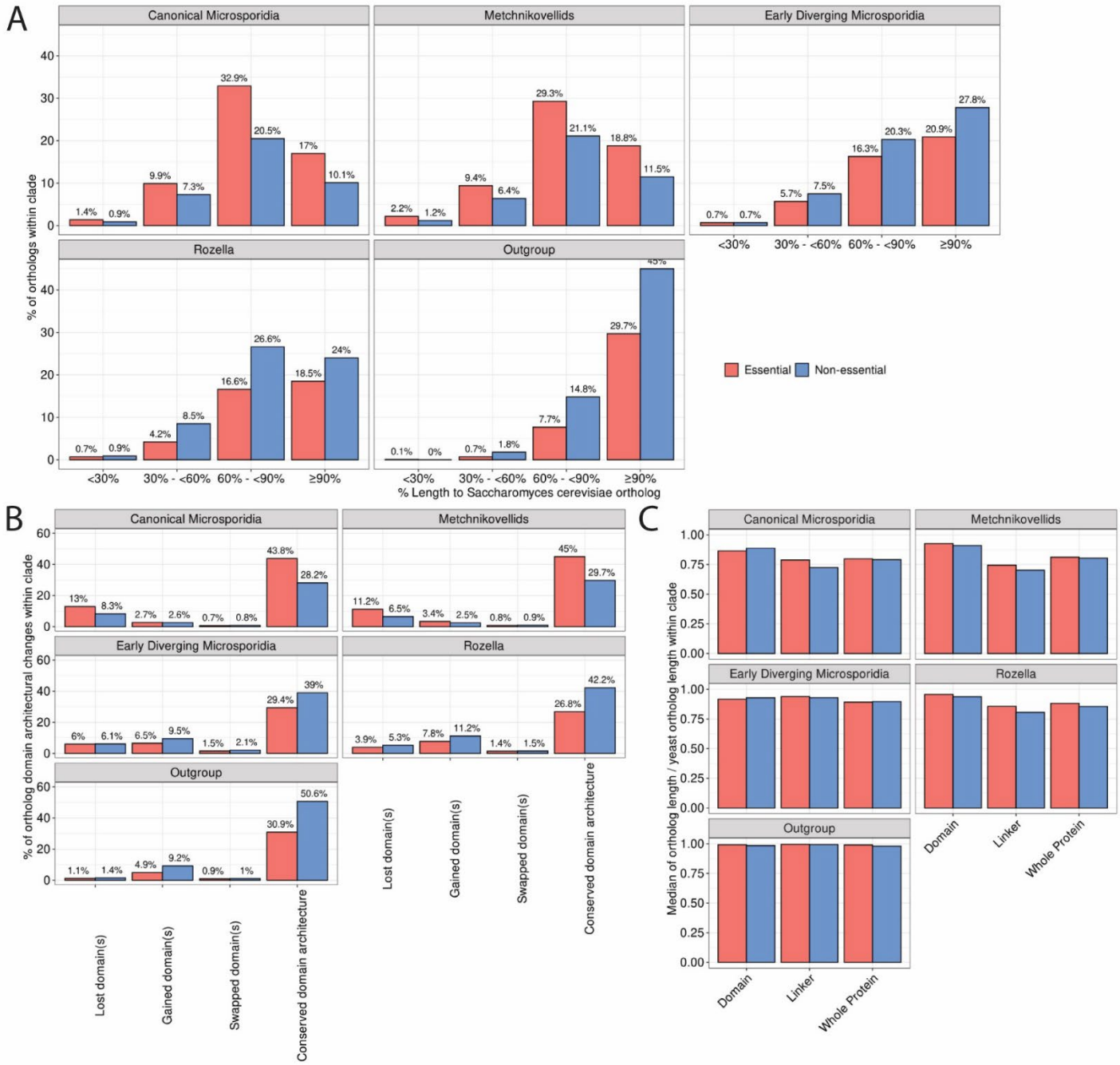


Fig. 2. Microsporidia proteins are shorter and have lost domains compared to their yeast orthologs. (A-C) Comparison of ortholog properties of the indicated groups of species stratified by yeast-essential (red) and non-essential (blue) orthologs. (A) Distribution of protein length relative to yeast ortholog. (B) Number of indicated domain architectural change events. Orthologs with multiple domain architecture change events (ex: domain loss + domain swap) are counted toward all categories. (C) Relative length of full-length protein, domain, and linker, relative to yeast ortholog.

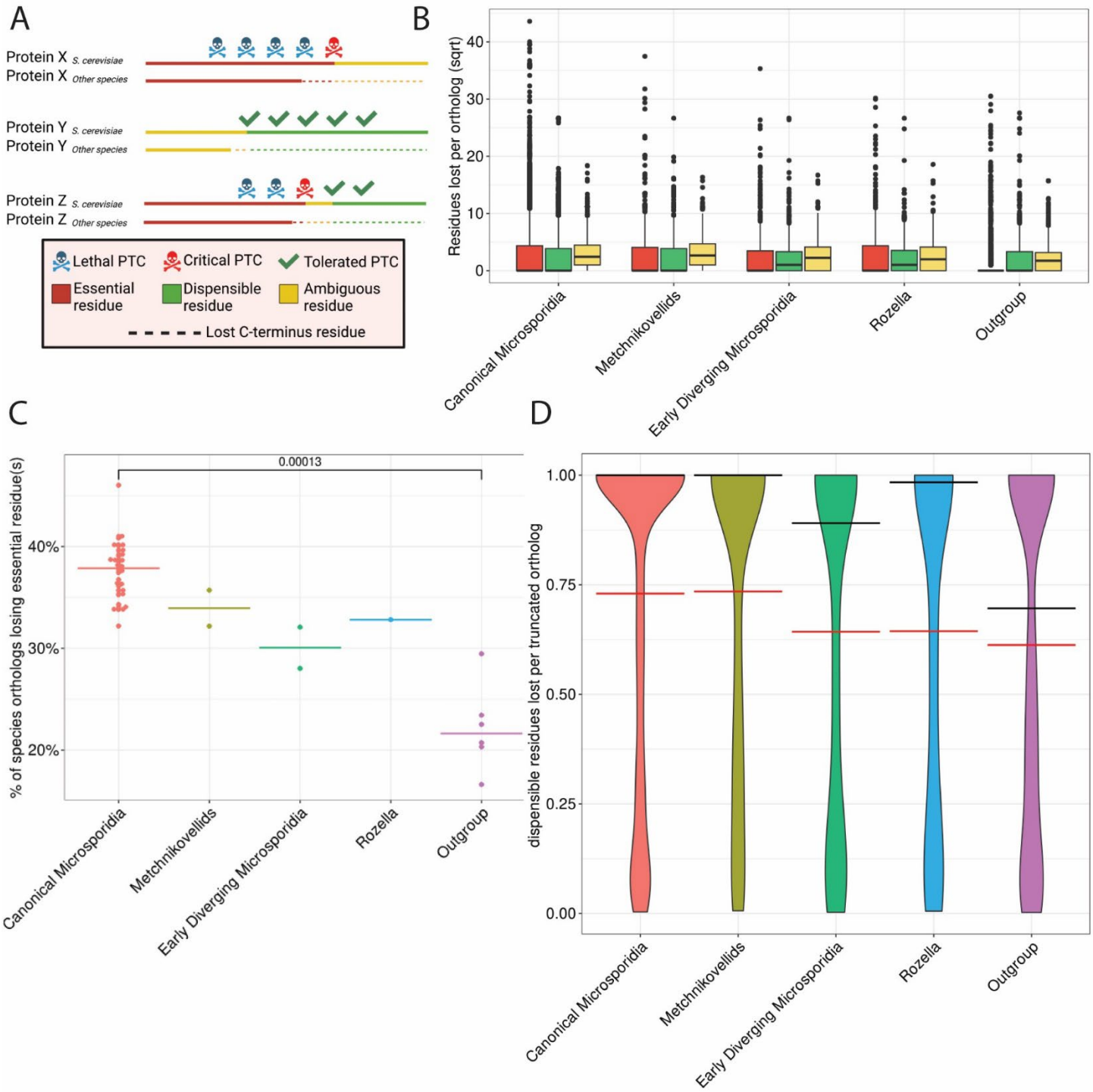


FIG. 3. Microsporidia orthologs have lost amino acids that are essential in yeast. (A) Cartoon schematic for amino acid essentiality classifications in essential yeast genes, using premature stop codon (PTC) tolerance data (Sadhu et al. 2018). Based on a PTC being either lethal or tolerated, residues were classified as dispensable, essential, or ambiguous. (B) Box-plot showing distributions of lost amino acid types in yeast-essential orthologs for each species group. (C) Percentage of proteins from each species that are shorter than their yeast ortholog that have lost any amino acids classified as essential. Each species is a data point and the average between different species groups is shown. P-value determined using Mann-Whitney test. (D) Distribution of the percentage of dispensable residues lost in each yeast ortholog. Data is presented as a violin plot and the median in each group is shown with a black line and the mean with a red line.

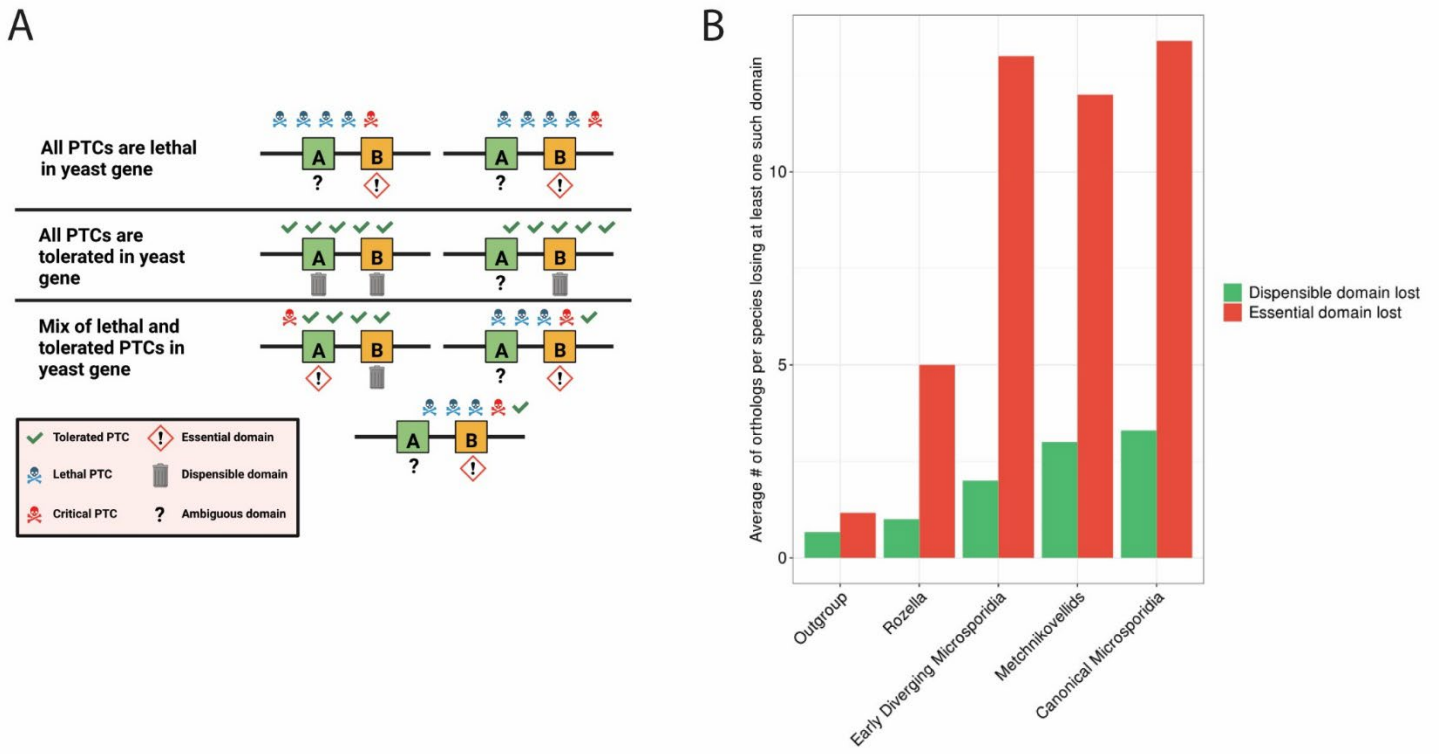


FIG. 4. Microsporidia orthologs have lost domains that are essential in yeast. (A) Cartoon schematic of domain classification in essential yest genes with PTC tolerance (Sadhu et al. 2018). (B) Composition of lost domains in yeast-essential orthologs, displayed as different groups showing the average number of proteins that have lost domains classified as either essential or dispensable.

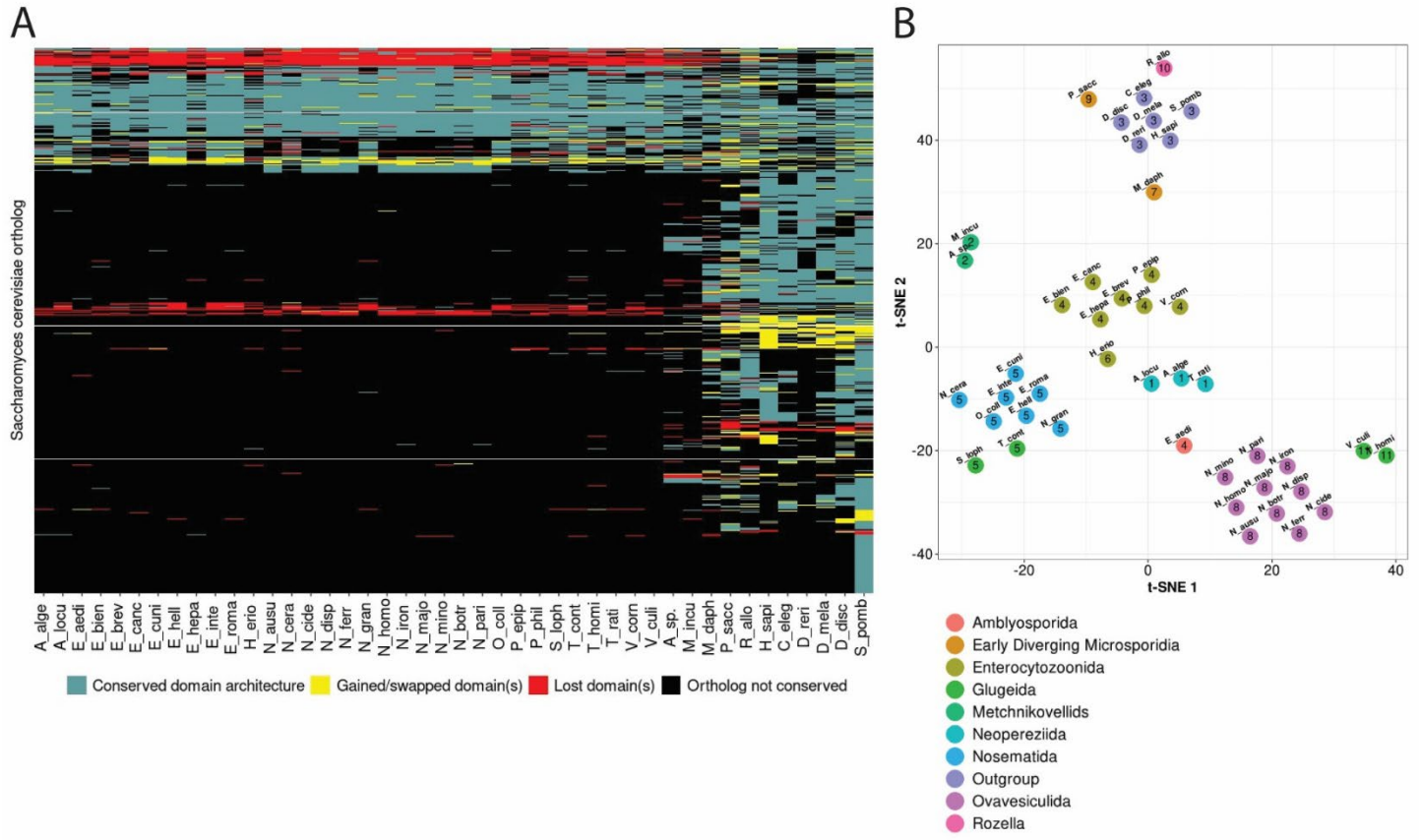


Fig. 5. Domain architectural changes in microsporidia proteins relative to yeast occurred early in evolution, in patterns consistent with phylogeny. (A) Domain architecture conservation status of microsporidia orthologs to yeast. Rows are yeast orthologs found in at least 1 microsporidia species, and columns are orthologs from a single microsporidia species. Yeast orthologs 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 plot of species by composition of lost Pfam clans in orthologs. Points are colored by phylogenetic clade of species (Bojko et al. 2022) and numbered by cluster assigned through k-medoids clustering of species by lost clan content.

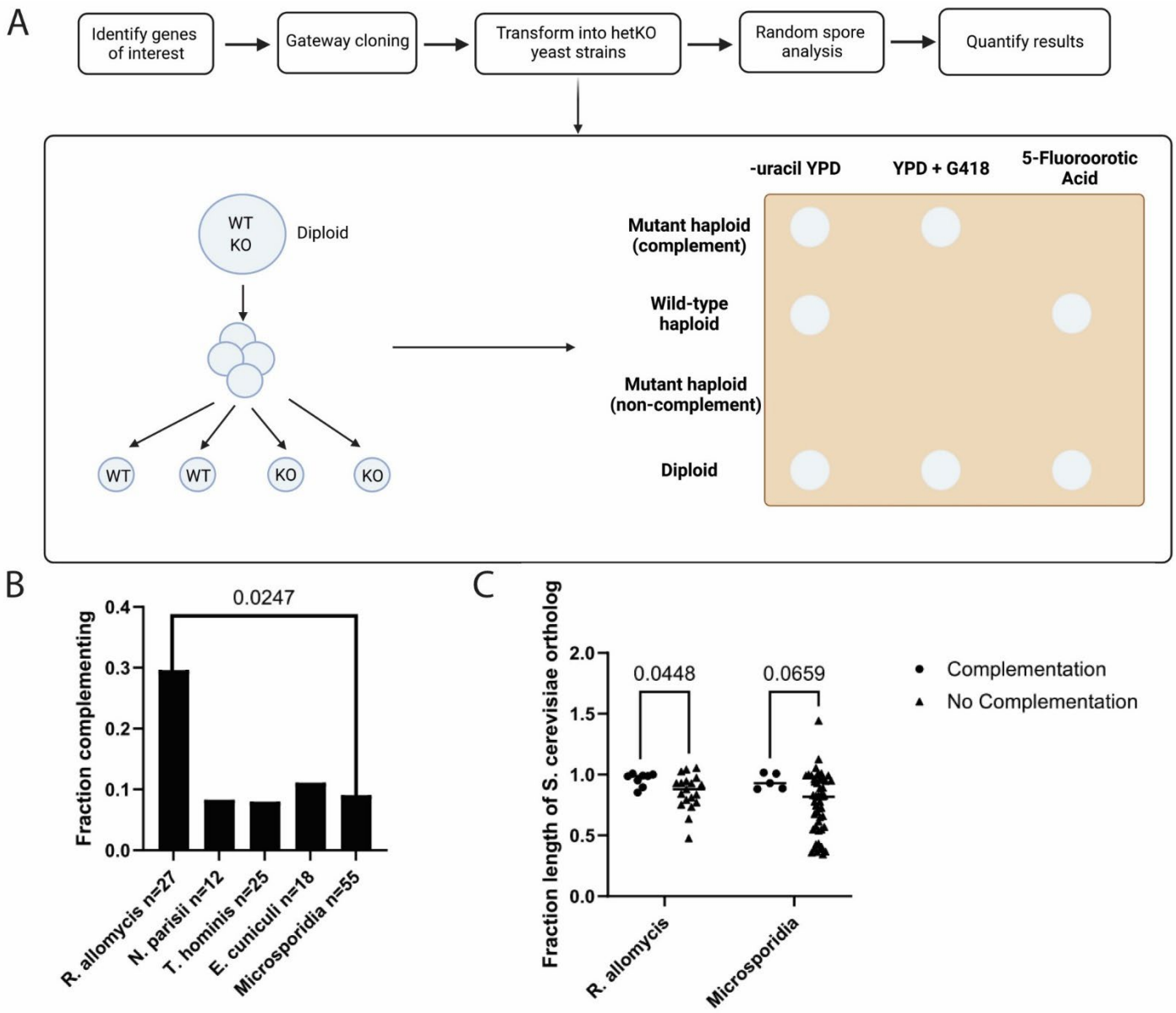


Fig. 6. Microsporidia orthologs are less likely to complement yeast deletions than *R. allomycis* and more reduced orthologs are less likely to complement. A. Schematic of random spore analysis assay to determine complementation. (B) Percentage of essential yeast genes complemented by the corresponding ortholog from the indicated species. P-value was determined by two-sided Fisher's exact test. (C) The ability of *R. allomycis* or microsporidia proteins to complement their yeast ortholog deletions compared to the length of each protein divided by length of the yeast ortholog. The p-values were determined by Mann-Whitney test.

594 **Supplementary figures and tables**

595

596 **Supplementary table S1. List of genome assemblies used in this study.**

597 **Supplementary table S2. Length, essentiality, and identity of microsporidia orthologs of yeast proteins.**

598 **Supplementary table S3. Conservation of domain architecture.**

599 **Supplementary table S4. Domains and linker lengths.**

600 **Supplementary table S5. Comparison of PTC data of microsporidia orthologs.**

601 **Supplementary table S6. Classification of lost domains in yeast-essential orthologs.**

602 **Supplementary table S7. Identity of domains conserved or lost in yeast orthologs.**

603 **Supplementary table S8. Results of random spore analysis assays.**

604 **Supplementary table S9. Microsporidia ortholog complementation of essential yeast proteins.**

605 **Supplementary table S10. Previously reported results of complementation of yeast mutants with**
606 **microsporidia genes.**

607 **Supplementary table S11. Oligos and sequences of genes used in this study.**

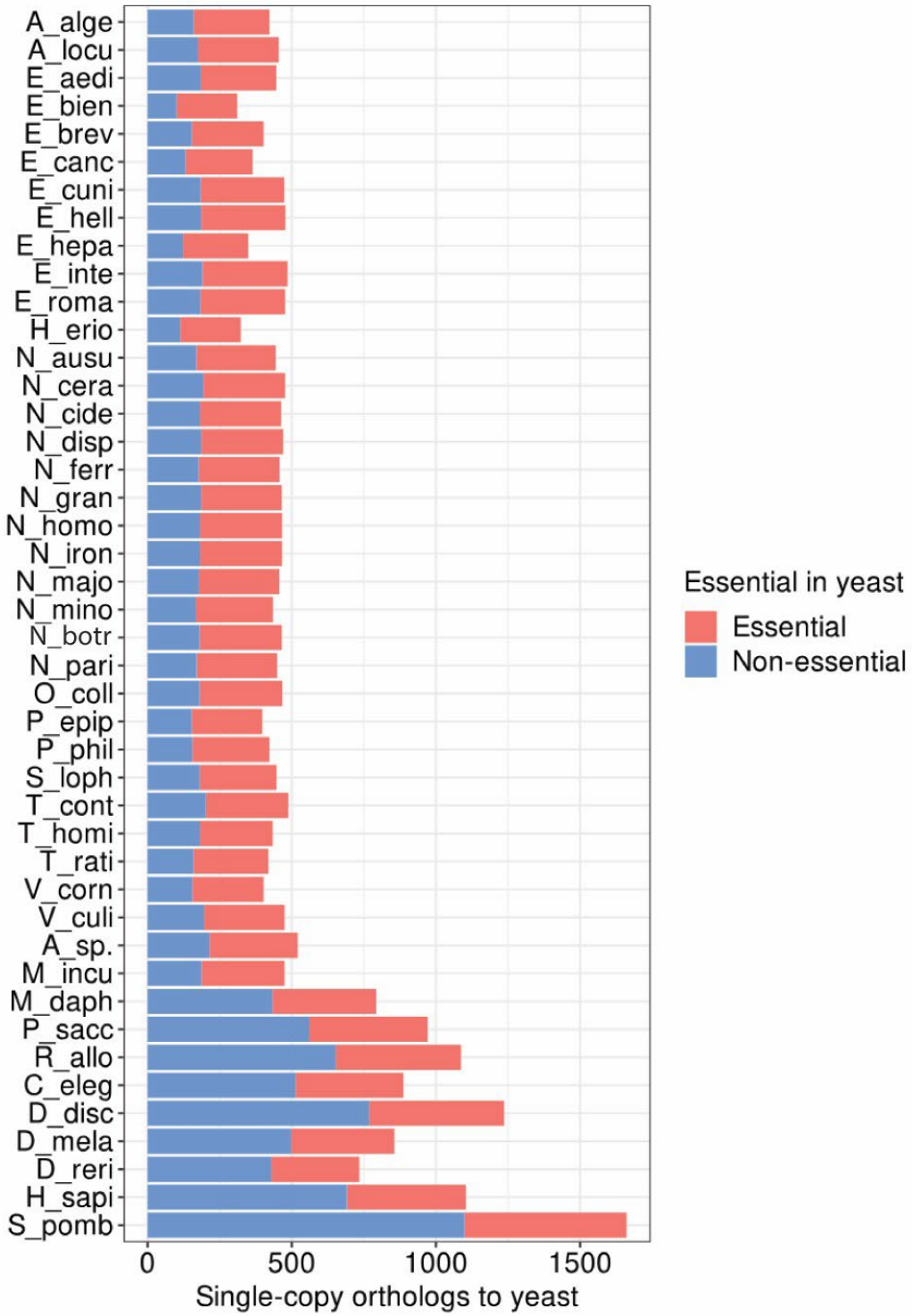
608

609

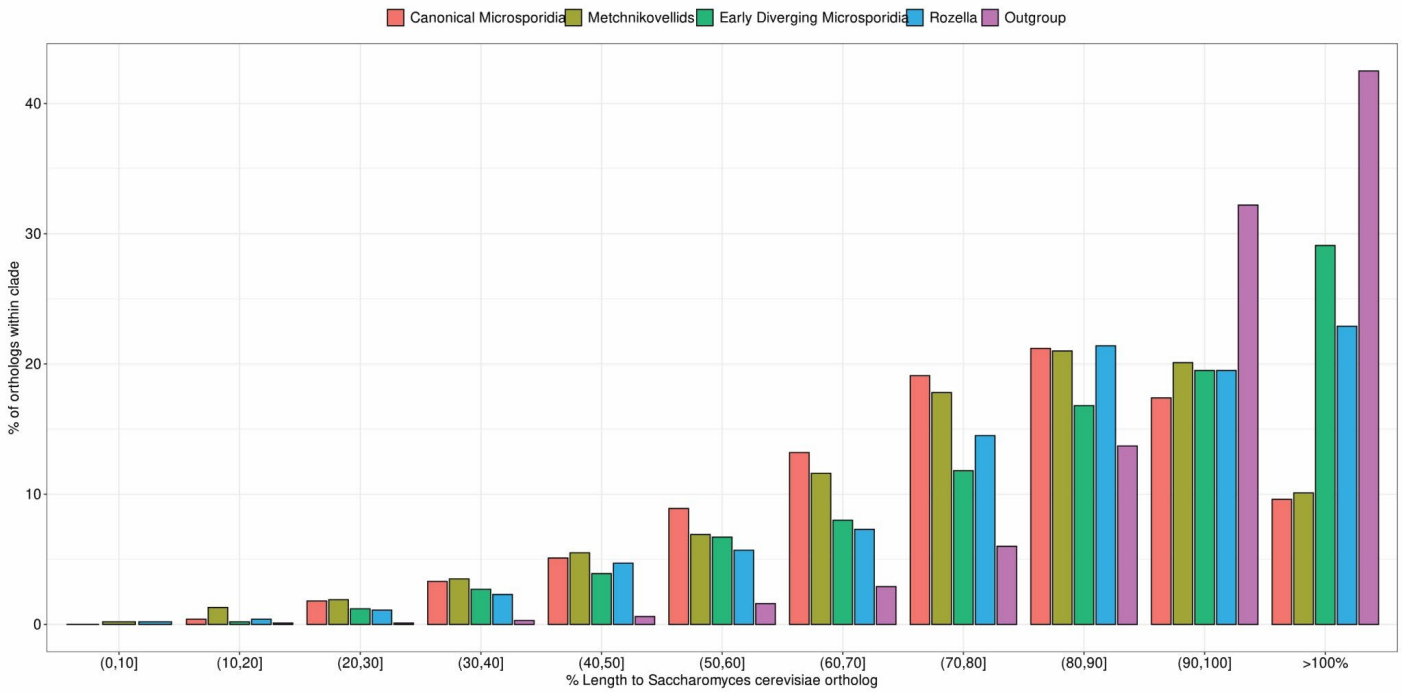
610

611

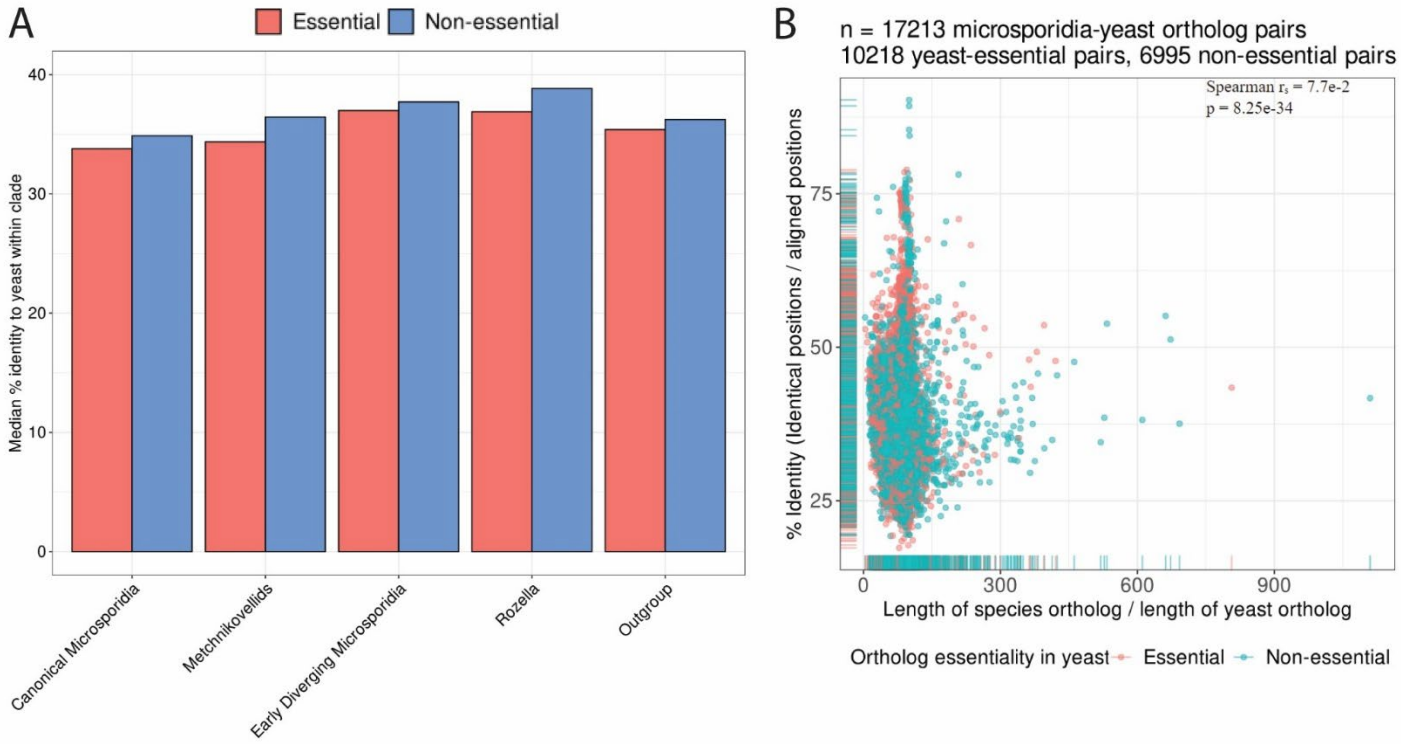
612



Supplementary fig. S1. Number of single-copy orthologs from each species that are essential or non-essential in yeast. The number of single-copy orthologs that each species has compared to yeast is shown, with orthologs of essential and non-essential genes in yeast shown according to the legend at the right.

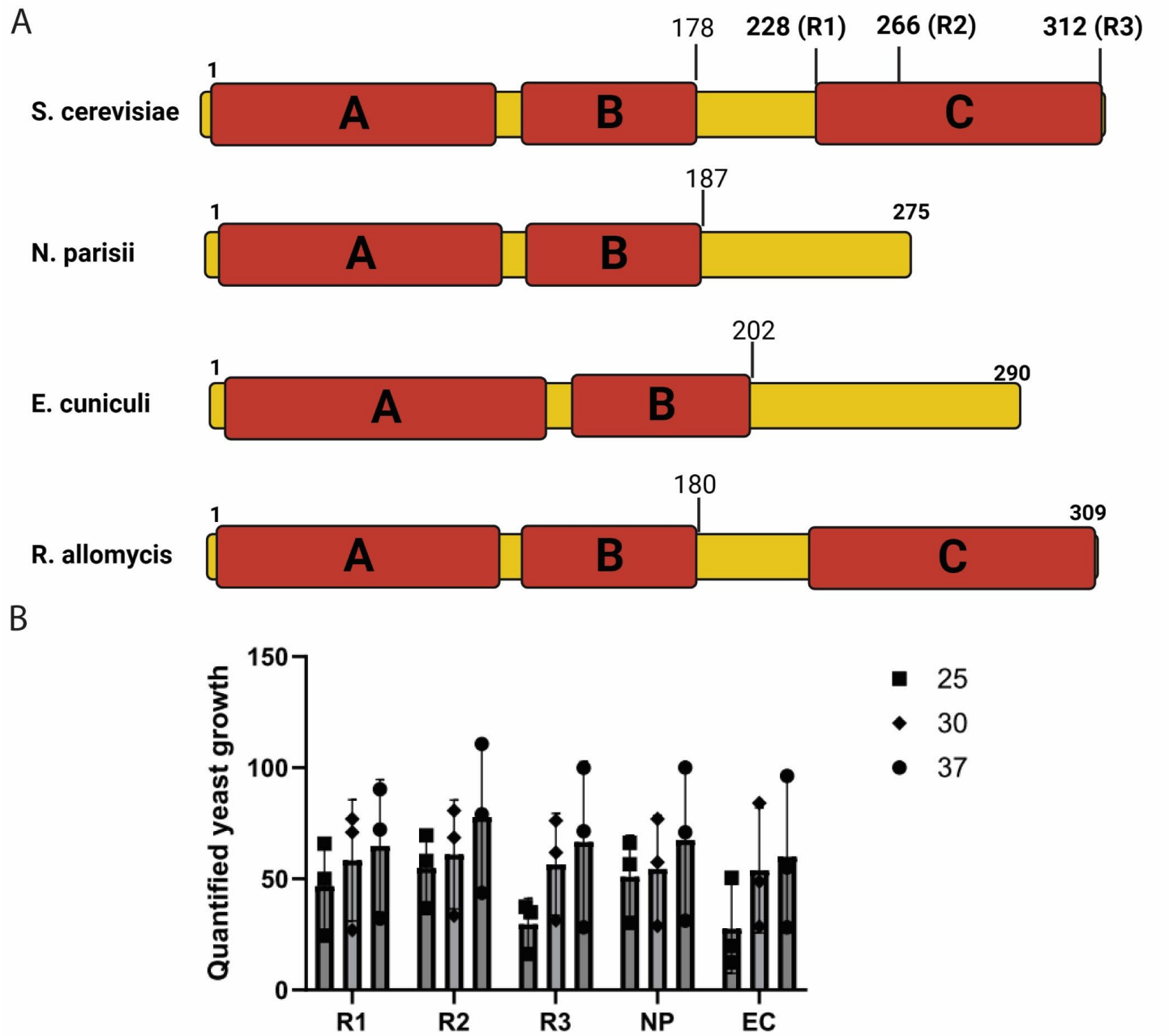


Supplementary fig. S2. Length distribution of orthologs compared to yeast. Distribution of the relative ortholog lengths relative to yeast for each species group.



Supplementary fig. S3. Sequence identity of orthologs. (A) Distribution of sequence identity relative to yeast ortholog of the indicated groups of species stratified by yeast-essential and non-essential orthologs. (B) Relative lengths of microsporidia orthologs to yeast compared to their percent identities.

632



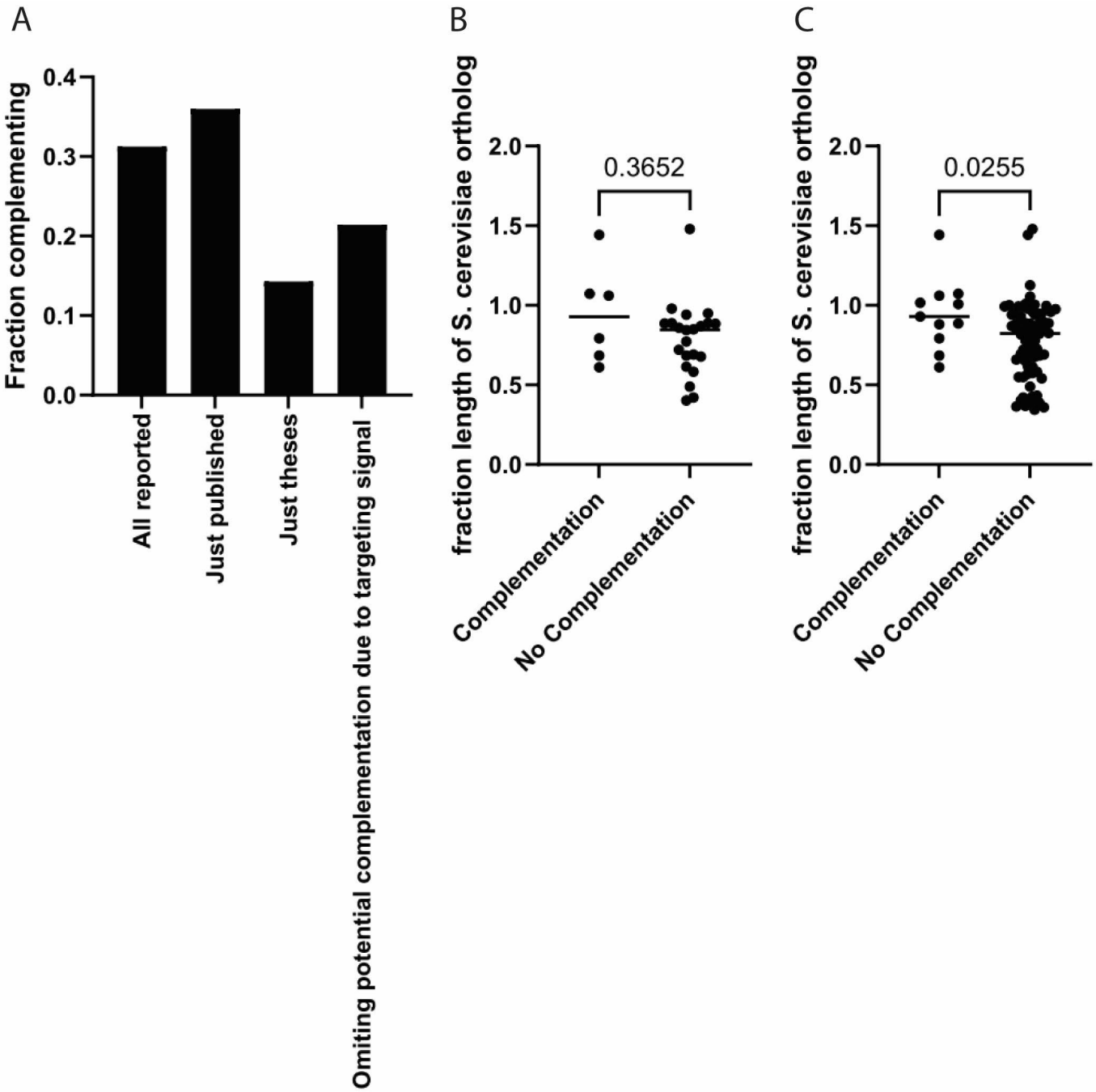
633

634

635 **Supplementary fig. S4. *S. cerevisiae* RPPO truncations rescue to similar extent as microsporidia**
636 **orthologs.** (A) Domain diagram of *S. cerevisiae* and microsporidia RPP0 proteins. Domain identities as given:
637 A (Large ribosomal subunit protein uL10), B (Large ribosomal subunit protein uL10-like, insertion domain), C
638 (60s Acidic ribosomal protein). (B) Various RPP0 containing-plasmids were transformed into a hetKO deletion
639 of RPPO. These strains were dilution plated at the indicated temperature and colony size was quantified. All
640 statistical comparisons between strains at each temperature have p- values > 0.5 determined using Two-way
641 ANOVA.

642

643



644

645 **Supplementary fig. S5. Previously reported complementation by microsporidia orthologs.** (A) Fraction of
646 microsporidia proteins that complement deletion strains of their yeast orthologs. (B-C) Ratio of microsporidia/
647 yeast length for each protein that could complement or not complement. (B) Literature reported
648 complementation, omitting potential complementation due to targeting signals. (C) Combined literature
649 complementation (fig. S5B) and complementation from this study (fig. 6C). The p-values were determined by
650 Mann-Whitney test.

651

652 **References**

653 Barandun J, Hunziker M, Vossbrinck CR, Klinge S. 2019. Evolutionary compaction and adaptation visualized by
654 the structure of the dormant microsporidian ribosome. *Nat Microbiol* 4:1798–1804.

655 Barrera A, Alastruey-Izquierdo A, Martín MJ, Cuesta I, Vizcaíno JA. 2014. Analysis of the Protein Domain and
656 Domain Architecture Content in Fungi and Its Application in the Search of New Antifungal Targets. *PLOS*
657 *Computational Biology* 10:e1003733.

658 Bass D, Czech L, Williams BAP, Berney C, Dunthorn M, Mahé F, Torruella G, Stentiford GD, Williams TA. 2018.
659 Clarifying the Relationships between Microsporidia and Cryptomycota. *Journal of Eukaryotic*
660 *Microbiology* 65:773–782.

661 Bojko J, Reinke AW, Stentiford GD, Williams B, Rogers MSJ, Bass D. 2022. Microsporidia: a new taxonomic,
662 evolutionary, and ecological synthesis. *Trends in Parasitology* 38:642–659.

663 Brocchieri L, Karlin S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*
664 33:3390–3400.

665 Corsaro D. 2022. Insights into Microsporidia Evolution from Early Diverging Microsporidia. In: Weiss LM, Reinke
666 AW, editors. Microsporidia: Current Advances in Biology. Experientia Supplementum. Cham: Springer
667 International Publishing. p. 71–90. Available from: https://doi.org/10.1007/978-3-030-93306-7_3

668 Davis MW, Jorgensen EM. 2022. ApE, A Plasmid Editor: A Freely Available DNA Manipulation and Visualization
669 Program. *Frontiers in Bioinformatics* [Internet] 2. Available from:
670 https://www.frontiersin.org/article/10.3389/fbinf.2022.818619

671 Dean P, Sendra KM, Williams TA, Watson AK, Major P, Nakjang S, Kozhevnikova E, Goldberg AV, Kunji ERS,
672 Hirt RP, et al. 2018. Transporter gene acquisition and innovation in the evolution of Microsporidia
673 intracellular parasites. *Nat Commun* 9:1709.

674 Desjardins CA, Sanscrainte ND, Goldberg JM, Heiman D, Young S, Zeng Q, Madhani HD, Becnel JJ, Cuomo
675 CA. 2015. Contrasting host-pathogen interactions and genome evolution in two generalist and specialist
676 microsporidian pathogens of mosquitoes. *Nat Commun* 6:7121.

677 Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Computational Biology* 7:e1002195.

678 Ehrenbolger K, Jespersen N, Sharma H, Sokolova YY, Tokarev YS, Vossbrinck CR, Barandun J. 2020.
679 Differences in structure and hibernation mechanism highlight diversification of the microsporidian
680 ribosome. *PLOS Biology* 18:e3000958.

681 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome*
682 *Biology* 20:238.

683 Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. *BMC*
684 *Bioinformatics* 12:326.

685 Freibert S-A, Goldberg AV, Hacker C, Molik S, Dean P, Williams TA, Nakjang S, Long S, Sendra K, Bill E, et al.
686 2017. Evolutionary conservation and in vitro reconstitution of microsporidian iron-sulfur cluster
687 biosynthesis. *Nat Commun* 8:13932.

688 Galindo LJ, Torruella G, Moreira D, Timpano H, Paskerova G, Smirnov A, Nassonova E, López-García P. 2018.
689 Evolutionary Genomics of Metchnikovella incurvata (Metchnikovellidae): An Early Branching
690 Microsporidium. *Genome Biol Evol* 10:2736–2748.

691 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, et
692 al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418:387–391.

693 Giaever G, Nislow C. 2014. The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics*
694 197:451–465.

695 Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R, Embley TM.
696 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. *Nature*
697 452:624–628.

698 Haag KL, James TY, Pombert J-F, Larsson R, Schaer TMM, Refardt D, Ebert D. 2014. Evolution of a
699 morphological novelty occurred before genome compaction in a lineage of extreme parasites. *PNAS*
700 111:15480–15485.

701 Hausmann S, Schwer B, Shuman S. 2004. An Encephalitozoon cuniculi Ortholog of the RNA Polymerase II
702 Carboxyl-Terminal Domain (CTD) Serine Phosphatase Fcp1. *Biochemistry* 43:7111–7120.

703 Hausmann S, Vivarès CP, Shuman S. 2002. Characterization of the mRNA Capping Apparatus of the
704 Microsporidian Parasite Encephalitozoon cuniculi*. *Journal of Biological Chemistry* 277:96–103.

705 Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, Harris SR, Weinmaier T, Markert S, Becher
706 D, et al. 2012. The genome of the obligate intracellular parasite Trachipleistophora hominis: new insights
707 into microsporidian genome dynamics and reductive evolution. *PLoS Pathog.* 8:e1002979.

708 James TY, Pelin A, Bonen L, Ahrendt S, Sain D, Corradi N, Stajich JE. 2013. Shared Signatures of Parasitism
709 and Phylogenomics Unite Cryptomycota and Microsporidia. *Current Biology* 23:1548–1553.

710 Jespersen N, Ehrenbolger K, Winiger RR, Svedberg D, Vossbrinck CR, Barandun J. 2022. Structure of the
711 reduced microsporidian proteasome bound by PI31-like peptides in dormant spores. *Nat Commun* 13:1–
712 14.

713 Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web
714 interface. *Nucleic Acids Res* 36:W5–W9.

715 Kachroo AH, Laurent JM, Akhmetov A, Szilagyi-Jones M, McWhite CD, Zhao A, Marcotte EM. 2017. Systematic
716 bacterialization of yeast genes identifies a near-universally swappable pathway.Barkai N, editor. *eLife*
717 6:e25093.

718 Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. 2015. Systematic humanization of
719 yeast genes reveals conserved functions and genetic modularity. *Science* 348:921–925.

720 Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P,
721 Wincker P, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite
722 Encephalitozoon cuniculi. *Nature* 414:450–453.

723 Keeling P. 2009. Five Questions about Microsporidia. *PLOS Pathogens* 5:e1000489.

724 Keeling PJ, Fast NM. 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu*
725 *Rev Microbiol* 56:93–116.

726 Kofoed M, Milbury KL, Chiang JH, Sinha S, Ben-Aroya S, Giaever G, Nislow C, Hieter P, Stirling PC. 2015. An
727 Updated Collection of Sequence Barcoded Temperature-Sensitive Alleles of Yeast Essential Genes. *G3*
728 *Genes|Genomes|Genetics* 5:1879–1887.

729 Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28:2520–
730 2522.

731 Krijthe JH. 2015. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation.
732 Available from: https://github.com/jkrijthe/Rtsne

733 Lai H-Y, Yu Y-H, Jhou Y-T, Liao C-W, Leu J-Y. 2023. Multiple intermolecular interactions facilitate rapid evolution
734 of essential genes. *Nat Ecol Evol* 7:745–755.

735 Laurent JM, Garge RK, Teufel AI, Wilke CO, Kachroo AH, Marcotte EM. 2020. Humanization of yeast genes with
736 multiple human orthologs reveals functional divergence between paralogs. *PLOS Biology* 18:e3000627.

737 Lewis TE, Sillitoe I, Lees JG. 2019. cath-resolve-hits: a new tool that resolves domain matches suspiciously
738 quickly. *Bioinformatics* 35:1766–1767.

739 Li Z, Vizeacoumar FJ, Bahr S, Li J, Warringer J, Vizeacoumar FS, Min R, VanderSluis B, Bellay J, DeVit M, et
740 al. 2011. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat*
741 *Biotechnol* 29:361–367.

742 Lind AE, Lewis WH, Spang A, Guy L, Embley TM, Ettema TJG. 2018. Genomes of two archaeal endosymbionts
743 show convergent adaptations to an intracellular lifestyle. *ISME J* 12:2655–2667.

744 Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation
745 and sequence length. *BMC Evolutionary Biology* 2:20.

746 Maaten L van der, Hinton G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9:2579–
747 2605.

748 Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2022. cluster: Cluster Analysis Basics and Extensions.
749 Available from: https://CRAN.R-project.org/package=cluster

750 Mascarenhas dos Santos AC, Julian AT, Pombert J-F. 2022. The Rad9–Rad1–Hus1 DNA Repair Clamp is
751 Found in Microsporidia. *Genome Biology and Evolution* 14:evac053.

752 Mathur V, Kolísko M, Hehenberger E, Irwin NAT, Leander BS, Kristmundsson Á, Freeman MA, Keeling PJ. 2019.
753 Multiple Independent Origins of Apicomplexan-Like Parasites. *Current Biology* 29:2936-2941.e5.

754 Melnikov SV, Rivera KD, Ostapenko D, Makarenko A, Sanscrainte ND, Becnel JJ, Solomon MJ, Texier C, Pappin
755 DJ, Soll D. 2018. Error-prone protein synthesis in parasites with the smallest eukaryotic genome. *Proc*
756 *Natl Acad Sci U S A* 115:E6245–E6253.

757 Mikhailov KV, Simdyanov TG, Aleoshin VV. 2017. Genomic Survey of a Hyperparasitic Microsporidian
758 Amphiamblys sp. (Metchnikovellidae). *Genome Biology and Evolution* 9:454–467.

759 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj
760 S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research*
761 49:D412–D419.

762 Moran NA. 2002. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell* 108:583–586.

763 Murareanu BM, Sukhdeo R, Qu R, Jiang J, Reinke AW, Weiss LM. 2021. Generation of a Microsporidia Species
764 Attribute Database and Analysis of the Extensive Ecological and Phenotypic Diversity of Microsporidia.
765 *mBio* 12:e01490-21.

766 Nakjang S, Williams TA, Heinz E, Watson AK, Foster PG, Sendra KM, Heaps SE, Hirt RP, Martin Embley T.
767 2013. Reduction and expansion in microsporidian genome evolution: new insights from comparative
768 genomics. *Genome Biol Evol* 5:2285–2303.

769 Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid
770 sequence of two proteins. *Journal of Molecular Biology* 48:443–453.

771 Nicholson D, Salamina M, Panek J, Helena-Bueno K, Brown CR, Hirt RP, Ranson NA, Melnikov SV. 2022.
772 Adaptation to genome decay in the structure of the smallest eukaryotic ribosome. *Nat Commun* 13:591.

773 Petropavlovskiy AA, Tauro MG, Lajoie P, Duennwald ML. 2020. A Quantitative Imaging-Based Protocol for Yeast
774 Growth and Survival on Agar Plates. *STAR Protocols* 1:100182.

775 Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et
776 al. 2012. The Pfam protein families database. *Nucleic Acids Research* 40:D290–D301.

777 Quandt CA, Beaudet D, Corsaro D, Walochnik J, Michel R, Corradi N, James TY. 2017. The genome of an
778 intranuclear parasite, Paramicrosporidium saccamoebae, reveals alternative adaptations to obligate
779 intracellular parasitism.Rokas A, editor. *eLife* 6:e29594.

780 R Development Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria:
781 R Foundation for Statistical Computing Available from: http://www.R-project.org

782 Reece-Hoyes JS, Walhout AJM. 2018. High-Efficiency Yeast Transformation. *Cold Spring Harb Protoc*
783 2018:pdb.prot094995.

784 Reinke AW, Troemel ER. 2015. The Development of Genetic Modification Techniques in Intracellular Parasites
785 and Potential Applications to Microsporidia. *PLoS Pathog* 11:e1005283.

786 Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal*
787 *of Computational and Applied Mathematics* 20:53–65.

788 Sadhu MJ, Bloom JS, Day L, Siegel JJ, Kosuri S, Kruglyak L. 2018. Highly parallel genome variant engineering
789 with CRISPR-Cas9. *Nat Genet* 50:510–514.

790 Santos C, Ballesta JPG. 1995. The Highly Conserved Protein P0 Carboxyl End Is Essential for Ribosome Activity
791 Only in the Absence of Proteins P1 and P2 (∗). *Journal of Biological Chemistry* 270:20608–20614.

792 Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*
793 9:671–675.

794 Slamovits CH, Burri L, Keeling PJ. 2006. Characterization of a Divergent Sec61β Gene in Microsporidia. *Journal*
795 *of Molecular Biology* 359:1196–1202.

796 Slyusarev GS, Starunov VV, Bondarenko AS, Zorina NA, Bondarenko NI. 2020. Extreme Genome and Nervous
797 System Streamlining in the Invertebrate Parasite Intoshia variabili. *Current Biology* 30:1292-1298.e3.

798 Soria PS, McGary KL, Rokas A. 2014. Functional Divergence for Every Paralog. *Molecular Biology and Evolution*
799 31:984–992.

800 Svedberg D, Winiger RR, Berg A, Sharma H, Tellgren-Roth C, Debrunner-Vossbrinck BA, Vossbrinck CR,
801 Barandun J. 2023. Functional annotation of a divergent genome using sequence and structure-based
802 homology. :2023.08.27.554996. Available from:
803 https://www.biorxiv.org/content/10.1101/2023.08.27.554996v2

804 Texier C, Brosson D, El Alaoui H, Méténier G, Vivarès CP. 2013. Post-genomics of microsporidia, with emphasis
805 on a model of minimal eukaryotic proteome: a review. *Folia Parasitologica* 52:15–22.

806 Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ. 2012. Mathematical modeling and comparison of protein
807 size distribution in different plant, animal, fungal and microbial species reveals a negative correlation
808 between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC*
809 *Research Notes* 5:85.

810 Upadhya R, Zhang HS, Weiss LM. 2006. System for Expression of Microsporidian Methionine Amino Peptidase
811 Type 2 (MetAP2) in the Yeast Saccharomyces cerevisiae. *Antimicrobial Agents and Chemotherapy*
812 50:3389–3395.

813 Wadi L, El Jarkass HT, Tran TD, Islah N, Luallen RJ, Reinke AW. 2023. Genomic and phenotypic evolution of
814 nematode-infecting microsporidia. *PLOS Pathogens* 19:e1011510.

815 Wadi L, Reinke AW. 2020. Evolution of microsporidia: An extremely successful group of eukaryotic intracellular
816 parasites. *PLoS Pathog* 16:e1008276.

817 Waller RF, Jabbour C, Chan NC, Celik N, Likić VA, Mulhern TD, Lithgow T. 2009. Evidence of a Reduced and
818 Modified Mitochondrial Protein Import Apparatus in Microsporidian Mitosomes. *Eukaryotic Cell* 8:19–26.

819 Wang D, Hsieh M, Li W-H. 2005. A General Tendency for Conservation of Protein Length Across Eukaryotic
820 Kingdoms. *Molecular Biology and Evolution* 22:142–147.

821 Wang M, Kurland CG, Caetano-Anollés G. 2011. Reductive evolution of proteomes and protein structures.
822 *Proceedings of the National Academy of Sciences* 108:11954–11958.

823 Warringer J, Blomberg A. 2006. Evolutionary constraints on yeast protein size. *BMC Evolutionary Biology* 6:61.

824

825

826

827

828