The *Ordospora colligata* Genome: Evolution of Extreme Reduction in Microsporidia and Host-To-Parasite Horizontal Gene Transfer

Jean-François Pombert, a,b Karen Luisa Haag, c Shadi Beidas, b Dieter Ebert, b, d Patrick J. Keeling a

Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada; Department of Biological and Chemical Sciences, Illinois Institute of Technology, Chicago, Illinois, USA; Department of Genetics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil; Department of Evolutionary Biology, Zoological Institute, University of Basel, Basel, Switzerland d

D.E. and P.J.K. are co-senior authors.

ABSTRACT Microsporidia are a group of obligate intracellular parasites that are best known for their unique infection mechanism and their unparalleled levels of genomic reduction and compaction. We sequenced the genome of *Ordospora colligata*, a gut parasite of the microcrustacean *Daphnia* sp. and the closest known relative to the microsporidia characterized by the most extreme genomic reduction, the model genus *Encephalitozoon*. We found that the *O. colligata* genome is as compact as those of *Encephalitozoon* spp., featuring few introns and a similar complement of about 2,000 genes, altogether showing that the extreme genomic reduction took place before the origin of *Encephalitozoon* spp. and their adaptation to vertebrate hosts. We also found that the *O. colligata* genome has acquired by horizontal transfer from its animal host a septin that is structurally analogous to septin 7, a protein that plays a major role in the endocytosis-based invasion mechanism of the fungal pathogen *Candida albicans*. Microsporidian invasion is most often characterized by injection through a projectile tube, but microsporidia are also known to invade cells by inducing endocytosis. Given the function of septins in other systems, we hypothesize that the acquired septin could help *O. colligata* induce its uptake by mimicking host receptors.

IMPORTANCE The smallest known eukaryotic genomes are found in members of the *Encephalitozoon* genus of microsporidian parasites. Their extreme compaction, however, is not characteristic of the group, whose genomes can differ by an order of magnitude. The processes and evolutionary forces that led the *Encephalitozoon* genomes to shed so much of their ancestral baggage are unclear. We sequenced the genome of *Ordospora colligata*, a parasite of the water flea *Daphnia* sp. and the closest known relative of *Encephalitozoon* species, and show that this extreme reduction predated the split between the two lineages. We also found that *O. colligata* has acquired a septin gene by host-to-parasite horizontal transfer and predicted that the encoded protein folds like a septin 7, which plays a major role in endocytosis. We hypothesize that this acquisition could help *O. colligata* parasitize its hosts by facilitating endocytic infection, a mechanism that occurs in microsporidia but that is not yet well understood.
have been so drastically altered: reduction is sometimes linked to their obligate intracellular parasitic lifestyle (10, 11), but only vaguely, and this is not obviously consistent with the 10-fold variability in other microsporidia.

The closest known relative of members of the Encephalitozoonidae is Ordospora colligata, which infects the microcrustacean Daphnia magna (12). Infections are typically located in the anterior part of the host’s midgut, where they can reach very high intensities, with nearly every gut epithelium cell being infected (13). The consequences of infection are not as severe as those seen with some microsporidia, but the reproductive success of infected females is reduced by 20%, and they die earlier than uninfected controls (14). Infected hosts release transmission-stage parasites with their feces, and the free-floating spores are then ingested by other filter-feeding Daphnia spp. The parasite has been reported from D. magna populations in Europe and the Middle East (13, 15).

As the sister to the Encephalitozoon genome, the O. colligata genome might offer some important clues to the origin and evolution of this model for genomic reduction and compaction. Here we describe the complete sequence of the nuclear genome of O. colligata OC4 and compare it to those of its relatives from the lineage Encephalitozoonidae. We show that the structures and contents of the O. colligata Oc4 and Encephalitozoon genomes are remarkably similar, with only a few distinct chromosomal reorganizations and very limited differences in gene content, showing that the extreme reduction characterizing Encephalitozoon genomes significantly preceded the origin of the genus. The most surprising difference between the two is that the O. colligata genome has acquired a Daphnia-derived septin by horizontal gene transfer (HGT). This protein is structurally analogous to septin 7 and retains transmembrane domains, suggesting that it could be located at the spore surface. In the fungal pathogen Candida albicans, the invasion mechanism is mediated by proteins called invasins that interact with the host septin 7 to induce endocytosis of the parasite by the host endothelial microfilaments (16). The presence of a Daphnia-derived septin 7 in O. colligata could facilitate its attachment to epithelial cells of Daphnia spp. by binding directly to its N-cadherin-like surface receptors and increase the likelihood of infection either simply by maintaining proximity to the host or more directly by triggering the host’s endocytosis mechanism.

RESULTS

Genome structure of O. colligata. The assembled O. colligata OC4 genome resulted in a total of 2,290,528 bp of unique sequence distributed in 15 contigs (627× average coverage). The 12 largest O. colligata single-copy contigs are structurally similarly to those of the 11 Encephalitozoon chromosomes, although there is evidence of some large transpositions such as the equivalents of chromosomes V and IX split and attached to other loci (Fig. 1). In contrast, the remaining three small contigs (5,456 to 16,597 bp) are repeated up to four times in the genome based on their respective coverage data and correspond to subtelomeric regions found in the Encephalitozoon species that could not be linked unambiguously by PCR.

One of the main structural differences between the O. colligata and Encephalitozoon genomes pertains to the telomeric regions. In the Encephalitozoon genomes, the rRNA operons are present in the subtelomeric regions of each chromosome, for a total of at least 22 copies (6–9). In contrast, the rRNA operons are present only four times in the O. colligata genome, indicating that its subtelomeric regions are structured differently from those in Encephalitozoon species. The sharp increase in GC richness in the Encephalitozoon subtelomeric regions compared with their chromosome cores (Fig. 1) was not observed in O. colligata; however, the cores themselves display the same arcing G+C% pattern as the Encephalitozoon species (Fig. 1) (7, 9), peaking in the central portions. At 38.2%, the overall G+C content of the O. colligata genome is lower than that of Encephalitozoon species and, perhaps not coincidentally, closest to that of the invertebrate pathogen E. romaleae (Table 1).

Evolution of gene content in O. colligata and Encephalitozoon spp. To first confirm the relationship between O. colligata and Encephalitozoon spp. that has previously been inferred from small-subunit (SSU) rRNA (17), we reconstructed the phylogeny based on 104 proteins that are shared between all microsporidian species investigated. This analysis robustly confirmed the positioning of O. colligata at the base of the Encephalitozoon species (see Fig. S1 in the supplemental material) (17), a phylogenetic affiliation that is also supported by gene content and metabolic profiling (see below). O. colligata is not closely related to Hamiltosporidium tvaerinminnensis (formerly called Octospora bayeri), which also infects Daphnia spp. (18) and, in some parts of their species range, coinfects the same host individual as O. colligata (19). Daphnia spp. are susceptible to many microsporidian species, which are widespread across the phylogenetic tree (17, 20, 21).

Annotating all open reading frames using homology to known proteins and positional orthology with E. cuniculi resulted in a total of 1,801 discrete protein-coding genes found in the O. colligata assembly. Overall, this makes the O. colligata genome only slightly less compact than those of its Encephalitozoon relatives, with very similar coding density (0.82 genes/kb versus 0.83 genes/kb in E. cuniculi), gene content, and lack of repeats and similar intron distribution (Table 1). Most of the genes that were identified are shared with Encephalitozoon species, even in those cases where Encephalitozoon spp. differ from other microsporidia. For example, O. colligata is incapable of endogenous RNA interference and lacks the Dicer and Argonaute proteins found in the 6-Mbp genomes of the microsporidian species Nosema ceranae, Spraguea lophii, and Trachipleistophora hominis (22–24). Unsurprisingly, the folate-related genes that were acquired by HGT in the Encephalitozoon lineage leading to E. hellem and E. romaleae (6) are also not found in O. colligata. However, the two ricin B-lectin domain-containing paralogs that are conserved across microsporidian species (22) and duplicated in tandem in the E. cuniculi strains (from ECU08_1700 to ECU08_1730) (7, 9) are also absent from the O. colligata genome. Instead, a single open reading frame (ORF) (M896_091670) that does not display any significant homology to genes encoding other proteins is located in the corresponding locus, suggesting that the ricin b-lectin paralogs were either amplified in Encephalitozoon spp. or perhaps lost or heavily modified and reduced in O. colligata.

Up to 95% of the O. colligata proteome is shared with the Encephalitozoon species (Fig. 2). In contrast, only 74% identity is found in Nosema species (a member of the sister group corresponding to the O. colligata/Encephalitozoon clade), with percentage decreasing rapidly as one moves further away in the tree, for an averaged pairwise proportion of 50% (highest, 100%; lowest,
16%). However, these low percentages do not necessarily mean that the genes are nonhomologous, since they may simply reflect the high rate of sequence divergence occurring in microsporidia. None of the *O. colligata* proteins that are absent from *Encephalitozoon* spp. have been found in other microsporidia, and, with the exception noted below, none have identifiable functions. These unique protein-coding genes are not restricted to subtelomeric regions; many are inserted within the cores of the *O. colligata* chromosome between genes arrayed in otherwise syntenic fashion with other *Encephalitozoon* spp. However, in 14 cases that included the M896_091670 gene inserted in lieu of the β-lectin genes, positional orthologs that are, however, dissimilar in sequences are found between the *O. colligata* and *Encephalitozoon* cores, suggesting that the corresponding genes may not be unique to *O. colligata* but may rather be divergent beyond recognition.

About 58% of the *O. colligata* proteome can be assigned putative functions, consistent with other microsporidia, for which 29% to 61% of the proteome has identifiable functions. The relative distributions of functions across metabolic pathways for identifiable proteins in microsporidia are similar across most species (Fig. 2; see also Table S1 in the supplemental material), with an elevated proportion of proteins involved in amino acid and carbohydrate metabolism in the genomes from basal *Rozella* and *Mitosporidium* species, which is congruent with their closer position-
are not silent; PTP2 displays very little similarity with its
tween the two genera. The nucleotide changes observed for PTP2
with polar tube protein 1 (PTP1; M896_060260) and polar tube
evolving genes between the two genera. This contrasts sharply
protein 2 (PTP2; M896_060250) stands out as one of the fastest-
absolute [K(JC)] rates and relative [K(JC)/Pi] rates, polar tube
spp. [K(JC)] (see Table S2 in the supplemental material). In both
domain located inside the core of the
peptide has been preserved, but the 8-amino-acid-long lysine-rich
proteins (26) is only barely recognizable in
lase (M896_060900; orthologous to ECU06_0910) involved in
chitinase (M896_021680; orthologous to ECU09_1320) involved
between the two genera with identifiable functions include an endo-
litozoon
Encephalitozoon
and
Encephalitozoon
tally in the gene-rich genomes of
sequenced microsporidian, 45% more than the second highest
nents, the latter of which represent the highest proportion in any
spp. or other microsporidia
. In all three cases, the overall G+C percentage is
about 10% lower than the content in the surrounding regions.
Two of the three segments (on contigs 3 and 8) are inserted within
blocks of genes that are otherwise syntenic with the Encephalitozoon
genomes, whereas the third (on contig 2) is located at one of the
junctions of a major intrachromosomal transposition involving
chromosome 2 (Fig. 3). We confirmed the assembly of all three
regions by PCR and Sanger sequencing, excluding potential as-
sembly errors.
The presumed transposition breakpoint is relatively gene poor,
but the other two regions are of comparable density to that of the
genome as a whole and encode several potential open reading
frames (Fig. 3). All but a few these ORFs are of unknown function.

| TABLE 1 General features of the O. colligata and Encephalitozoon genomes |
|-----------------|----------------|------------------|------------------|------------------|
| Feature         | O. colligata OC4 | E. cuniculi GB-M1 | E. intestinalis ATCC 50506 | E. hellem ATCC 50504 | E. romaleae SI-2008 |
| No. of chromosomes | 10              | 11               | 11               | 11               | 11               |
| Estimated total genome size (Mbp) | 3.0             | 2.9              | 2.3              | 2.5              | 2.5              |
| Assembled-genome size (Mbp) | 2.3             | 2.5              | 2.2              | 2.3              | 2.2              |
| Genome coverage (%) | 77              | 86               | 96               | 92               | 88               |
| G+C content (%) | 38.2            | 47               | 41.4             | 43.4             | 40.3             |
| Gene density (gene/kbp) | 0.62          | 0.63             | 0.91             | 0.89             | 0.84             |
| Mean gene length (bp) | 1041           | 1041             | 999              | 1018             | 1061             |
| Mean intergenic length (bp) | 176            | 166              | 100              | 106              | 130              |
| No. of SSU-LSU RNA genes | 4              | 22               | 22               | 22^a             | 22               |
| No. of 5S rRNA genes | 3              | 3                | 3                | 3                | 3                |
| No. of ncRNAs^d | 10             | NA               | 14               | 11               | NA               |
| No. of tRNAs | 46             | 46               | 46               | 46               | 46               |
| No. of RNA introns (sizes in bp) | 2 (11, 42)   | 2 (12, 41)       | 2 (12, 41)       | 2 (12, 41)       | 2 (12, 41)       |
| No. of splic. introns (size range in bp) | 30 (23–77) | 36 (23–76)       | 36 (23–76)       | 36 (23–76)       | 36 (23–76)       |
| Predicted no. of ORFs | 1820          | 2010             | 1944             | 1928             | 1835             |

^a The E. cuniculi and E. romaleae values are from Pombert et al. (6). The E. intestinalis and E. hellem values are from the 2014 accession updates.
^b The numbers of chromosomes for O. colligata were estimated based on the chromosomal reorganizations observed with the Encephalitozoon species.
^c The numbers of small-subunit and large-subunit (SSU-LSU) rRNA genes were inferred based on their overall coverage relative to other genes.
^d NA, the ncRNAs in E. cuniculi and E. romaleae were not assessed. They are likely similar in number and content to those found in O. colligata, E. intestinalis, and E. hellem.
**FIG 2.** Pairwise distribution and metabolic profiling of the microsporidian proteome. The phylogenetic positions in the cladogram are derived from our phylogenetic inferences (see Fig. S1 in the supplemental material). Nodes recovered in all of the bootstrap replicates are indicated by asterisks, whereas values indicate levels of support for the corresponding nodes. The early-diverging microsporidian *Mitosporidium* and the Cryptomycota *Rozella* spp. were used as outgroups in this analysis. Assembled-genome sizes are indicated in Mbp between each branch and the corresponding microsporidian taxa. In the adjacent heat map, light and dark colors indicate the percentages of proteins from the branching species that are shared with the species indicated on top. Darker colors indicate (Continued)
**FIG 3** O. colligata genomic loci that are absent from other microsporidia. The regions shown under the 20-kbp scale bars are color coded according to Fig. 1 and are drawn to scale, with the corresponding G+C content plotted underneath. Numbers between the chromosomal loci and G+C plots indicate the corresponding genomic regions shown under the 20-kbp scale bars are color coded according to Fig. 1

**DISCUSSION**

The *Encephalitozoon* genus has been held as a model for extreme genome reduction and compaction, and the patterns left by these processes over the evolution of the genus have been well studied (6, 8, 9, 28). Explanations for why these genomes are so compact are less clear, but the 10-fold variation in genome size in microsporidia suggests that it is not intrinsically related to intracellular parasitism. Here, we show that the extreme reduction in the *Encephalitozoon* genome actually predates the origin of the genus and is common to the *Daphnia*-infecting sister lineage represented by *O. colligata*. Indeed, in terms of gene density, content, and overall structure, the *O. colligata* genome is hardly distinguishable from those of *Encephalitozoon* spp. The impressive level of synteny observed between the two genera strongly argues against independent rounds of reductions, such that the genome of the last common ancestor of the *O. colligata* and *Encephalitozoon* species was almost assuredly similar in every aspect.

Of the differences in gene content we do describe, however, the septin 7 homologue stands out as being of particular functional and evolutionary interest. Obligate intracellular parasites have limited opportunities to exchange genetic material, given that their sheltered environment offers few opportunities to exchange genes with organisms other than the host. Turning a host’s proteome against itself is a strategy that has been used by both prokaryotic and eukaryotic pathogens throughout the course of evolution, although clear-cut cases of host-to-parasite HGT in eukaryotes are still rare in general (29, 30). Horizontal gene transfer has been shown in microsporidians, despite their genome-reductionist tendencies, and some of their newly acquired genes have obvious possible benefits with respect to infection or survival outside their host (31, 32). Only a single case of HGT from an animal host has been reported in microsporidia, however (6, 33), and here we show that a septin 7 in *O. colligata* not only is animal...
0.3 substitution per site

FIG 4 Maximum likelihood phylogenetic inferences of *O. colligata*’s septin 7 (M896_080490) and similar protein sequences. All protein sequences display similarty greater than the selected cutoff (*E* value cutoff 1E to 40; accession date, 20 August 2014) were retrieved from the NCBI NR database. The best ML derived but is apparently closely related to homologues in *D. pulex* and *D. magna* genomes. The directionality of this transfer is clear; no homologues are found in other fungi, and the transfer is likely recent as well. The horizontal transfer seems to have occurred after the split between the *O. colligata* and *Encephalitozoon* spp. and to have resulted in the acquisition of additional transmembrane domains since its origin. The alternative, acquisition by their common ancestor and then loss in the members of the *Encephalitozoonidae*, seems unlikely given their host range histories.

The function of the *O. colligata* septin is unknown, but other fungal pathogens offer intriguing possibilities related to cell division and compartmentalization (34). Intuitively, cell division proteins form valuable acquisition targets for vertically transmitted parasites to modulate the reproductive cycle of hosts. Here, however, the function of the HGT-acquired septin is unlikely to be involved in cell division: unlike the microsporidian *Daphnia* pathogens *Perezia diaphanosoma*, which can infect both gut and reproductive organs (12), *O. colligata* is a strictly horizontally transmitted gut epithelium pathogen that has never been encountered in reproductive organs. Thus, a sequested septin would not disrupt host reproduction. There is also no evidence that this HGT-acquired septin is secreted and, considering the evolutionary distance between the parasite and its hosts, it is unlikely that this protein would play a role in *O. colligata*’s own cell division. One could, however, imagine a function in compartmentalization, for example, inducing endocytosis by the host. *Encephalitozoon* species can invade their host by endocytosis (2, 3), and in the fungal pathogen *Candida albicans*, the endocytic invasion is initiated by proteins called invasins that interact with the host septin 7, a major effector (35), which initiates a molecular cascade, ultimately inducing the uptake of the pathogen (16). The septin acquired by *O. colligata* is structurally analogous to septin 7 and features additional transmembrane motifs, suggesting that it could be localized at the proteinaceous exosporum. An externally exposed septin could camouflage *O. colligata* and bypass the need for invasions by binding to cell surface *N*-caderhins of *Daphnia* spp., by recruiting other components of the host septin 2/6/7 complex, or by interacting directly with the microfilaments of the gut epithelium to facilitate entry into the host by endocytosis. Alternatively, a surface septin could also facilitate infection by simply helping to keep the parasite in close proximity to the host cell surface.

Unfortunately, little is known about the infection process in *O. colligata* (or in many other microsporidia for that matter), so more direct evidence is required to determine if *O. colligata* is capable of infection by self-induced endocytosis and, if so, if the HGT-acquired septin is surface localized and facilitates this process. There is no guarantee that the gene has remained functional since being acquired by HGT, and while the septin core appears to have been conserved in *O. colligata* (see Fig. S2 in the supplemental material), the long branch it displays in phylogenetic analyses (Fig. 4) suggests that it is evolving at a fast pace. In any case,
because *Encephalitozoon* species are capable of infecting their host by endocytosis (3) and yet lack the septin 7 found in *O. colligata*, we infer that this gene is not essential to the underlying mechanism. However, considering the number of diverse microsporidian species that infect *Daphnia* spp., the presence of a functional septin 7 in *O. colligata* could confer to it a competitive advantage over its parasitic relatives. The distribution of this gene in related taxa is also of interest for efforts to help determine how recent the HGT was: it may have originated relatively early and been lost in *Encephalitozoon* spp. or may be present within only a limited number of genotypes.

**MATERIALS AND METHODS**

**Tissue culture and DNA purification.** *O. colligata* isolate OC4 was cultivated in *Daphnia magna* clone ELK1-1 (England-LadyKirk-pond 1) (36) and used for further laboratory procedures. Approximately 1,000 female hosts infected with *O. colligata* OC4 were homogenized in 10 mM Tris-HCl (pH 7) and then filtered sequentially through a 40-μm-pore-size nylon mesh and an 8-μm-pore-size cellulose nitrate membrane adapted to a syringe. The filtrate was centrifuged at 4,000 rpm for 10 min and resuspended in 10 mM Tris-HCl (pH 7). The solution, containing spores and other tissue debris, was centrifuged in 60% Percoll (Sigma Aldrich) at 14,000 rpm for 5 min, and the pellet was washed 3 times with 10 mM Tris-HCl (pH 7) to obtain a clear spore solution. Spores were incubated with lysozyme (Sigma Aldrich) (2.5 mg/ml) at 37°C for 1 h to lysate contaminating bacteria. An additional step of contaminant cell lysis was performed by adding a lysis buffer (1% SDS; 2% Triton X-100; 1 mg/ml proteinase K; 10 mM Tris-HCl; 1 mM EDTA; 100 mM NaCl; pH 7.0) to the mixture and incubating it at 56°C for 1 h. The lysate was centrifuged at 14,000 rpm to recover the spores, which were treated with DNase I (Sigma Aldrich) at 37°C overnight to eliminate contaminating DNA. The enzyme was inactivated with EGTA (50 mM) at 95°C for 30 min, and the spores recovered by centrifugation were frozen and thawed several times. The...
O. colligata purified spores were used for DNA isolation using a DNeasy tissue-extracting kit (Qiagen). Before sequencing, genomic DNA was assessed for quality by Qubit fluorometric quantification (Life Technologies).

Sequencing. The O. colligata purified DNA Illumina 100-bp paired-end (PE) libraries (212-bp inserts, 43-bp average standard deviation, 40,096,994 PE reads, 8,019,398,800 bp total) were prepared using TrueSeq SBS V5 chemistry and sequenced by Fasteris (Geneva, Switzerland) on an Illumina HiSeq 2000 instrument. Reads were processed and the adapters removed with CASAVA pipeline version 1.8.1 (mean quality score, 33.90; SBS V5 chemistry and sequenced by Fasteris (Geneva, Switzerland) on an end (PE) libraries (212-bp inserts, 43-bp average standard deviation, January/February 2015 Volume 6 Issue 1 e02400-14 85.37% of all bases ≥ Q30). Read quality was further assessed with FASTQC (37).

Genome assembly. Paired-end reads were filtered by calculation of quality scores with Sickle 1.210 ([Biobinformatics Core, University of California, Davis https://github.com/najoshi/sickle]) and the filtered reads iteratively assembled de novo with Ray 2.0-r8 using odd k-mer values (from 19 to 31) on 32 Infiniband QDR-connected Intel Xeon nodes (256 Nehalem X5650 processing cores at 2.8 GHz). Microsporidian contigs were filtered from host and other low-level contaminants in the resulting datasets with BLAST homology searches (38) using EnsemblEukaryot genomes (BLASTN) and proteins (tBLASTN) as queries. Microsporidian contigs from each k-mer assembly were concatenated and merged with Consed (39) and then visually inspected for potential discrepancies between the various k-mer assemblies. Merged contigs were extended using our in silico chromosome walking approach (7) with the addSolexaReads.pl script from the Consed package. Ambiguous regions in the assemblies were amplified by PCR using flanking primers and then validated by Sanger sequencing. Final assemblies were verified by read mapping with Bowtie 1.0 (40) and visual inspection with TableT.1.13.07.31 (41).

Genome annotation. Transfer and ribosomal RNAs were positioned on the O. colligata contigs with tRNAscan-SE 1.3.1 (42) and RNAmmer 1.2 (43), respectively. Open reading frames were first positioned on the sequences with Artemis 16.0.0 (44) built-in tools. Start methionines were then refined using multiple-sequence alignments of orthologs with MAFFT 7.058 b (45) and the presence of CCC/GGG-like transcription signals as described by Peyretaillade et al. (46). Noncoding RNAs (ncRNAs) were positioned using a combination of BLASTN homology searches, syntenic information from EnsemblEukaryot genomes, and RFAM (47) searches as implemented in Artemis. Putative protein-coding gene functions were ascribed by homology searches against UniProt (48), Pfam (49) searches, and InterProScan 5 (50) analyses. Microsatellites were searched for with WebSat (51) using default parameters.

Protein analyses and 3D structure predictions. The presence of signal peptides in primary amino acid sequences was searched for with SignalP 4.1 (52). Transmembrane domain searches and hydrophobicity analyses were performed with TMHMM 2.0 (53), TopPred 2 (54), and TMpred (55) as implemented on the CBS (http://www.cbs.dtu.dk/services/), Institut Pasteur (http://mobyle.pasteur.fr/), and Expasy (http://www.expasy.org/) web portals, respectively. The M896_080490 primary amino acid sequence was converted from FASTA to PDB format and first modeled three-dimensionally with ROSETTA3 version 2014wk05 (56). This resulted in an unfolded chain of little value but containing several directional changes indicating a degree of secondary structure. The sequence was then run through Robetta (57), and two domains were determined via Ginzu analysis. These domains were queried against the NCBI nonredundant database (NR) using PSI-BLAST (58) to determine potential sequence homology to previously filed protein structures. Due to its higher confidence score, the second domain was analyzed with Web-based InterPro software (59). The highest-correlating gene ontologies (GOs) were analyzed with respect to previously predicted functions and correlated with the results of the Robetta analysis. The model was created with the most highly correlating base model via SWISS-MODEL (60) (Fig. 3) and correlated with its parent template by dot plot analysis as implemented in FFAS03 (61).

Phylogenetic analyses. For phylogenomic inferences, the microsporidian protein sequences were retrieved from the MicrosporidiaDB (62), SilkPathDB (http://silkpathdb.swu.edu.cn/silkpathdb/), and GenBank databases. Orthologous sequences were identified by BLASTP searches at an E value cutoff of 1E-20 using the O. colligata proteins as queries. Orthologs were aligned with MAFFT L-INS-I (45), and the ambiguous positions in the resulting alignments were filtered out with BMGE (63) using the default parameters. Maximum likelihood (ML) inference analyses were performed with PHYLML 3.0 (64) using the LG model of amino acid substitutions with four gamma categories. A total of 100 bootstrap replicate experiments were performed. Bootstrap replicates were generated with Seqboot and node percentages calculated with Consense from the PHYLIP 3.695 package (65). For horizontal gene transfer inference determinations, the M896_080490 orthologous and paralogous sequences were retrieved from the NCBI nonredundant (NR) database (accession date, 20 August 2014) using BLASTP searches with an E value cutoff of 1E-40. The retrieved sequences were aligned with MUSCLE (66), and the resulting alignment was filtered with BMGE using the default parameters. ML inference and bootstrap replicate experiments were performed as described above.

Pairwise distribution and metabolic profiling. For each species, local protein databases were generated using MAKEBLASTDB from the NCBI BLAST+ 2.2.28 package. The presence or absence of genes in comparisons between species was determined by evaluating pairwise BLASTP hits (E value cutoff, 1e-10) for all possible combinations. Metabolic profiles were inferred from InterProScan 5 (50) analyses performed on each protein data set; for each protein, the gene ontologies retrieved were filtered to remove duplicates and concatenated into higher hierarchies derived from the KEGG orthology pathways using custom Perl scripts.

Nucleotide diversity. The O. colligata protein-coding genes and their EnsemblEukaryot orthologs were aligned by codon comparisons performed with MACSE (67). For each alignment, the nucleotide diversity (Pi) between EnsemblEukaryot species and their divergence from O. colligata [K(I/C)] were inferred using the polymorphism and divergence in functional regions tool implemented in DnaSP 5/10/01 (68) with O. colligata as the outgroup.

Accession numbers. The O. colligata data were released in the NCBI database under BioProject PRJNA210314, BioSample SAMN02867507, and accession number JOKQ00000000.

SUPPLEMENTAL MATERIAL
Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.02400-14/-/DCSupplemental.

Figure S1, PDF file, 0.1 MB.
Figure S2, PNG file, 0.2 MB.
Table S1, XLSX file, 0.01 MB.
Table S2, XLSX file, 0.5 MB.

ACKNOWLEDGMENTS
This work was supported by a grant from the Canadian Institutes of Health Research to P.J.K. (MOP-42517; http://www.cihr-irsc.gc.ca/).

REFERENCES
6. Pombert J-F, Selman M, Burki F, Bardell FT, Farinelli L, Solter LF,


32. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and...


