



## Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth

Christina A. Cuomo, Christopher A. Desjardins, Malina A. Bakowski, et al.

*Genome Res.* published online July 18, 2012

Access the most recent version at doi:[10.1101/gr.142802.112](https://doi.org/10.1101/gr.142802.112)


---

<b>P&lt;P</b>	Published online July 18, 2012 in advance of the print journal.
<b>Accepted Manuscript</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
<b>Creative Commons License</b>	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <a href="http://genome.cshlp.org/site/misc/terms.xhtml">http://genome.cshlp.org/site/misc/terms.xhtml</a> ). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at <a href="http://creativecommons.org/licenses/by-nc/3.0/">http://creativecommons.org/licenses/by-nc/3.0/</a> .
<b>Email Alerting Service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .

---

CRISPR and RNAi Genetic Screening.  
Your new superpower.

LEARN MORE



CELLECTA

---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

# **Title: Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth**

**Authors:** Christina A. Cuomo<sup>1†\*</sup>, Christopher A. Desjardins<sup>1†</sup>, Malina A. Bakowski<sup>2†</sup>, Jonathan Goldberg<sup>1</sup>, Amy T. Ma<sup>2</sup>, James J. Becnel<sup>3</sup>, Elizabeth S. Didier<sup>4</sup>, Lin Fan<sup>1</sup>, David I. Heiman<sup>1</sup>, Joshua Z. Levin<sup>1</sup>, Sarah Young<sup>1</sup>, Qiandong Zeng<sup>1</sup>, Emily R. Troemel<sup>2\*</sup>

**Affiliations:** <sup>1</sup> The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142

<sup>2</sup> Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093

<sup>3</sup> USDA, ARS, CMAVE 1600 S.W. 23rd Drive Gainesville, FL 32608

<sup>4</sup> Division of Microbiology, Tulane National Primate Research Center, 18703 Three Rivers Road, Covington, LA 70433 USA

†These authors made equivalent contributions

\*To whom correspondence should be addressed. E-mail: [cuomo@broadinstitute.org](mailto:cuomo@broadinstitute.org) (CAC)  
[etroemel@ucsd.edu](mailto:etroemel@ucsd.edu) (ERT)

**Running title:** Evolutionary strategies of microsporidia.

**Keywords:** microsporidia, obligate intracellular pathogen, *Nematocida*, *N. parisii*, *C. elegans*, Retinoblastoma, hexokinase, nucleoside transport, RNA-sequence, comparative genomics

## **Abstract:**

Microsporidia comprise a large phylum of obligate intracellular eukaryotes that are fungal-related parasites responsible for widespread disease, and here we address questions about microsporidia biology and evolution. We sequenced three microsporidian genomes from two species, *Nematocida parisii* and *Nematocida* sp1, which are natural pathogens of *Caenorhabditis* nematodes and provide model systems for studying microsporidian pathogenesis. We performed deep-sequencing of transcripts from a timecourse of *N. parisii* infection. Examination of pathogen gene expression revealed compact transcripts and a dramatic takeover of host cells by *Nematocida*. We performed phylogenomic analyses of *Nematocida* and other microsporidian genomes to refine microsporidian phylogeny and identify evolutionary events of gene loss, acquisition, and modification. In particular, we found that all microsporidia lost the tumor suppressor gene Retinoblastoma, which we speculate could accelerate the parasite cell cycle and increase mutation rate. We also found that microsporidia acquired transporters that could import nucleosides to fuel rapid growth. In addition, microsporidian hexokinases gained secretion signal sequences, and in a functional assay these were sufficient to export proteins out of the cell; thus hexokinase may be targeted into the host cell to reprogram it toward biosynthesis. Similar molecular changes appear during formation of cancer cells and may be evolutionary strategies adopted independently by microsporidia to proliferate rapidly within host cells. Finally, analysis of genome polymorphisms revealed evidence for a sexual cycle that may provide genetic diversity to alleviate problems caused by clonal growth. Together these events may explain the emergence and success of these diverse intracellular parasites.

Supplemental material is available for this article. The sequence data from this study have been deposited in Genbank ([www.ncbi.nlm.nih.gov/genbank](http://www.ncbi.nlm.nih.gov/genbank)) under accession no. AEFF01000000 for *N. parisii* ERTm1, AE0001000000 for *N. parisii* ERTm3, and AERB01000000 for *Nematocida* sp1 ERTm2. SNP data were submitted to dbSNP ([www.ncbi.nlm.nih.gov/projects/SNP/](http://www.ncbi.nlm.nih.gov/projects/SNP/)) and scheduled for Build 138 release for ERTm2 (handle: BROAD-GENOMEBIO, sbid 1057048); release schedule of ERTm1 and ERTm3 SNPs is pending. Living spores of *Nematocida* sp1 ERTm2 have been deposited in the American Type Culture Collection (ATCC) Protistology Collection as PRA-371.

## Introduction:

There are countless species of obligate intracellular microbes, which by definition are completely dependent on intracellular resources from their hosts. These microbes include pathogens as well as symbionts, which are thought to be the ancestors of intracellular organelles like mitochondria (Keeling 2011). The strategies used by obligate microbes to thrive within host cells remain enigmatic, despite obligate intracellular pathogens being responsible for significant medical and agricultural problems. For example, about 10% of college students in the United States (U.S.) are infected with the poorly understood pathogen *Chlamydia trachomatis*, which is an obligate intracellular bacterium and a leading cause of genital infections and blindness (James et al. 2008). In addition to prokaryotic pathogens, eukaryotic obligate intracellular pathogens also cause severe morbidity and mortality in a variety of hosts. For example, Microsporidia comprise a phylum with over 1,200 species of parasites that can infect and kill animals from virtually every phylum, with few treatments available (Keeling and Fast 2002; Williams 2009). Specifically, microsporidia cause opportunistic infections in AIDS patients, as well as organ transplant recipients, malnourished children, and the elderly (Didier and Weiss 2011). Microsporidia also cause agricultural disease: 50-94% of honeybees in the U.S. are infected with the microsporidian *Nosema ceranae*, which has been implicated in honeybee colony collapse disorder (Johnson et al. 2009; Troemel 2011).

Microsporidian genomes encode a greatly reduced metabolic potential compared to other eukaryotes (Keeling and Corradi 2011), yet these microbes can still undergo dramatic proliferation within host cells. This rapid growth likely places a substantial metabolic burden on the host to generate building blocks such as nucleotides, amino acids and lipids, which are

needed by the replicating parasite cells. How microsporidia, or other obligate intracellular microbes, direct the host to make these biosynthetic factors and how they gain preferential access to them is not well understood.

Although microsporidia appear to be most closely related to fungi, it is debated whether they should be included within the fungal kingdom or excluded from it (James et al. 2006). A major challenge in phylogenetic assignment of the Microsporidia is that they have highly divergent sequences (Stiller and Hall 1999; Katinka et al. 2001), which could be due to rapid proliferation and/or a high mutation rate. Previous phylogenetic analyses have relied on comparisons of only one or a small number of genes and have suggested multiple placements of microsporidia relative to fungi (Edlind et al. 1996; Hirt et al. 1999; Keeling et al. 2000). Comparisons of many shared genes among microsporidia and fungi could help resolve these controversies regarding Microsporidia phylogeny.

Although all species of microsporidia are thought to share general features in their life cycle such as invasion using a polar tube, replication as multi-nucleate meronts, and differentiation into spores, the details of their life cycle can be complex and they vary from species to species. On a broader scale, it is still unclear whether microsporidia are haploid or diploid, and whether they undergo a mating and meiosis cycle. Previously sequenced microsporidian genomes are presumed to be haploid. However, some species of microsporidia are thought to undergo a sexual cycle (Becnel et al. 2005), and recent studies have indicated that gene order of the zygomycete mating locus is conserved in the microsporidia (Lee et al. 2008). This observation suggests that microsporidia could have a diploid stage, a mating cycle and may be true fungi. Answering

questions of ploidy, as well as questions about mating and recombination could provide insight into microsporidian speciation and pathogenic mechanisms, since mating appears to be important for the generation of diversity and virulence in other pathogenic fungi (Ni et al. 2011).

Recently, we discovered natural parasites of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*, and showed that they define a new microsporidian genus (*Nematocida*) that contains two species (*parisii* and sp1) (Troemel et al. 2008). As *C. elegans* is a well understood model organism, the *N. parisii*-*C. elegans* system provides a powerful model in which to decipher microsporidian biology. Microsporidia, like most obligate intracellular pathogens, are refractory to genetic analysis, and thus genome analysis provides the best current alternative to investigate questions of evolution and pathogenesis of the microsporidia. Here we sequence the genomes of three *Nematocida* strains and perform deep-sequencing expression analysis during distinct stages of infection. We compare these genomes to several other divergent microsporidia genomes, as well as other eukaryotic genomes, to better understand the evolutionary strategies that led to the emergence of this large phylum of obligate intracellular pathogens.

## **Results:**

### ***Nematocida* genome assembly and content**

We sequenced and assembled the genomes of three *Nematocida* strains (Table S1): two strains of *N. parisii* collected from wild-caught *C. elegans* in France (ERTm1 and ERTm3) and one of a divergent species termed *Nematocida* sp1 collected from wild-caught *C. briggsae* in India (ERTm2) (Troemel et al. 2008). Each strain was propagated in the standard N2 laboratory strain

of *C. elegans*. The assemblies of the two *N. parisii* strains are very similar in size, each totaling 4.1 Mb (Table 1), and contain ~95% of their sequence on 8 or 9 scaffolds of ~150 to ~900 kb (Table S2). As some scaffolds may be complete chromosomes, this suggests there are less than 8 chromosomes in *N. parisii*. These genomes are 99.8% identical, and 96% of ERTm1 can be uniquely aligned to ERTm3. Thus, the low divergence of these two isolates supports their previous designation as the same species.

The genome of *Nematocida* sp1 (ERTm2) is highly diverged from the *N. parisii* (ERTm1 and ERTm3) assemblies. Alignments of ERTm1 and ERTm2 cover only 1.7 Mb, with 68.3% average identity and 80.2% average similarity in protein-based alignments. Despite this divergence these genomes appear highly syntenic with a small number of intra-scaffold inversions (Figure S1).

The *Nematocida* sp1 (ERTm2) assembly is 4.6 Mb, 14% larger than the assemblies of the two *N. parisii* strains. This expansion can be explained in part by a larger amount of repetitive sequence (2.1 fold more in *Nematocida* sp1) (Table 1, Table S2), some of which shares sequence similarity with transposable elements (see Supp. Methods).

### **RNA-seq analysis of gene expression**

To determine the transcribed regions of the *N. parisii* (ERTm1) genome and define gene expression during its infection cycle, we performed strand-specific deep sequencing of RNA (RNA-seq) isolated at distinct stages of *N. parisii* infection in *C. elegans*. *N. parisii*, like other microsporidia, undergoes a complex infection cycle, with distinct stages of parasite morphology (Figure 1A). To produce a synchronous infection for RNA-seq analysis, we inoculated *C. elegans* with a large dose of *N. parisii* spores and quantified infection progression by DIC

microscopy (Figure 1B, 1D). We also visualized infection with fluorescence *in situ* hybridization (FISH) staining, which illustrates the rapid proliferation of *N. parisii* throughout the infection cycle within the host (Figure 1C). Based on these results, we isolated RNA from *C. elegans* at five timepoints (8, 16, 30, 40 and 64 hours post inoculation (hpi) with *N. parisii*; Figure 1D) to capture gene expression at distinct stages of infection. The early timepoint samples (8, 16 and 30 hpi) contain only pathogens in the replicative meront stage, while later timepoint samples (40 and 64 hpi) contain a mixture of pathogens in the meront and in the transmissible spore stage.

In addition to microscopy, RNA-seq and qRT-PCR analysis of a mixed pathogen and host RNA sample illustrates how rapidly this pathogen replicates during infection. The abundance of *N. parisii* reads comprised only a small fraction of the sample at early timepoints, but represented an increasing percentage of RNA as infection progressed, a finding which was confirmed with qRT-PCR (Table S4). At 8 hpi, only 0.26% of the total RNA-seq reads corresponded to the *N. parisii* genome, and this fraction increased until 40 hpi, when approximately 28% of the reads corresponded to *N. parisii* (Figure 1D, Table S4), despite the pathogen being restricted to the worm intestine. Based on the increase in number of RNA-seq reads we estimated the doubling time of *N. parisii* to be 3.3 h during the replicative meront stage of infection (between 16 and 30 hpi), and based on qRT-PCR to be 2.9 h (see Supp. Methods). These estimates are close to the 2.9 h doubling time of *S. pombe* in rich culture media (Fantes 1977), suggesting a finely tuned adaptation of this pathogen to intracellular growth conditions.

We predicted 2,661 genes in the *N. parisii* (ERTm1) genome, 2,546 of which had RNA-seq based evidence for expression during infection (Table 1). These numbers are similar in

magnitude to the gene count from other microsporidia genomes, e.g. *Encephalitozoon cuniculi* (1,996) (Katinka et al. 2001), *Encephalitozoon intestinalis* (1,833) (Corradi et al. 2010), and *N. ceranae* (2,060) (Cornman et al. 2009). Forty-three percent of *N. parisii* proteins can be assigned a putative function from PFAM domains, and twenty of the forty-three percent also could be assigned gene ontology (GO) terms (Table 1). This is less than other microsporidia; using the same methods, 66% of *E. cuniculi* genes could be assigned PFAM domains and 52% could be assigned both PFAM domains and GO terms. The large number of *Nematocida*-specific genes (see below) could account for some of this difference. We also identified 198 proteins (7.5% of the proteome) predicted to be secreted in ERTm1 (Table 1), which is similar to predictions for *E. cuniculi* (5.2%) and ascomycete fungi *S. pombe* (4.5%) and *A. nidulans* (8.6%).

No introns were predicted in protein-coding genes in any of the three genomes. Based on the RNA-seq data, we also did not observe any spliced mRNA transcripts in these genomes. Numerous RNA components of the spliceosome machinery were missing (e.g., U2 and U6 snRNAs, U2AF splicing factor) as well as the SRPK and PRP4 kinases that regulate splicing activity, further supporting the predicted lack of any spliced transcripts in *Nematocida*. Splicing machinery also appeared lost in the *Enterocytozoon bieneusi* and *Vittaforma corneae* clades, an absence previously noted in *E. bieneusi* (Akiyoshi et al. 2009). The remaining microsporidia in the phylogenetic analysis (see below) appear to have intact splicing machinery, suggesting multiple independent losses of splicing within microsporidia.

The *N. parisii* (ERTm1) genome is compact, as 72.8% of the sequence is predicted to be coding, with a mean distance between coding sequences of 418 bp. To determine the relationship of

these short intergenic distances to transcript structure, we utilized the RNA-seq data to predict untranslated regions (UTRs) in ERTm1 and produce the first large-scale annotation of microsporidian UTRs. Most genes are encoded by discrete transcripts with either unusually short UTRs or no UTRs at all (Supp. Text, Figure S2), in contrast to the large multigenic transcripts previously identified in *E. cuniculi* (Gill et al. 2010). Calculation of RNA-seq read coverage across five feature types (coding, 5' UTR, 3' UTR, antisense, and intergenic) showed that antisense transcription is approximately equivalent to intergenic transcription (Figure S3). This result suggests there is little antisense transcription in *Nematocida* despite the close proximity of coding sequences. The small size of noncoding intergenic regions likely constrains the size of UTRs in *N. parisii* to avoid overlap with coding sequences. Within intergenic regions we identified five candidate promoter elements, including a TATA box and a modified version of the CCC motif previously described in *N. ceranae* (Supp. Text, Figure S4) (Cornman et al. 2009).

We also sequenced RNA from purified *N. parisii* spores and identified spore-specific transcripts by comparing these expression data to the infection timecourse data (Table S6). *N. parisii* spore transcripts are predominantly composed of ribosomal RNA with very little poly(A) RNA (see Supp. Text). We identified the genes most highly expressed overall in spores (Table S7), which include several transmembrane domain containing genes. These genes could encode structural proteins, or could encode receptor proteins that recognize cues that trigger firing of the polar tube, the initiating event in microsporidia infection.

### **Heterozygosity and loss of heterozygosity in *Nematocida* sp1 genome**

Analysis of the sequence data suggests that *Nematocida* sp1 (ERTm2) is diploid, highly heterozygous, and has undergone loss of heterozygosity (LOH) in regions of some scaffolds. A total of 42,175 heterozygous single nucleotide polymorphism (SNP) positions were identified in ERTm2 using Illumina sequencing. The reference and alternate allele are each supported by roughly equal proportions of the reads (the median allele balance at SNP positions is 0.56), in support of a diploid model. By examining the distribution of SNPs across the genome, we found that most of the assembly is heterozygous, but that scaffolds 1, 2, and 4 have each undergone LOH as evidenced by a single large homozygous region on each scaffold (Figure 2A). These homozygous regions include only 0.5% of the SNPs and comprise 21% of the genome, with an average of one SNP every 4,497 bases. The heterozygous regions include 99.5% of the SNPs and comprise 79% of the genome (examining scaffolds at least 30 kb in length), with an average of one SNP every 82 bases. SNPs are found in coding regions nearly as much as expected; 63.1% of SNPs are in coding regions, and coding regions cover 64.4% of the ERTm2 sequence. Most (69.2%) of the SNPs in coding regions result in synonymous amino acid substitutions. We also analyzed the 454 sequence reads for *N. parisii* and found that both ERTm1 and ERTm3 appear to be diploid and heterozygous, although there is a lower rate of polymorphism for these strains compared to *N. sp1* ERTm2 (Figure 2B).

The heterozygous and LOH regions of *Nematocida* sp1 (ERTm2) appear stable over several generations of being transmitted in the *C. elegans* host. The original ERTm2 DNA sample used for sequencing was derived from a pooling of many infected samples, which were prepared in parallel from a small number of infected donor animals (see Supp. Methods). Next, we serially passaged ERTm2 in *C. elegans* for 14 weeks and then isolated genomic DNA for sequencing.

The polymorphism pattern in this passaged isolate was nearly identical to the original isolate with no additional LOH regions (Figure 2A), indicating that ERTm2 was not undergoing rapid LOH during laboratory passage in the *C. elegans* host, and thus recombination is uncommon under these conditions.

Because all three *Nematocida* strains appeared to be diploid, suggesting that microsporidia may have a mating and a meiotic cell cycle, we identified potential orthologs of characterized mating and meiosis genes from other fungi (Table S8, S9). Of a set of 29 proteins suggested as a core meiotic set (Malik et al. 2008) mostly involved in chromosome cohesion and recombination, 20 orthologs are present in the *Nematocida* genomes. This pattern of conservation is very similar to that found in *E. cuniculi* and *N. ceranae*; one exception is the presence of Rec8 meiotic centromere cohesin found only in *Nematocida* species. All microsporidia appear to be missing several components (Apc1, Apc4, Apc5, and Cdc26) of the anaphase promoting complex. In summary, *Nematocida* genomes encode a few key meiosis genes not found in other microsporidian genomes (Table S8, S9). Perhaps these meiosis genes have been retained in *Nematocida* due to the shorter period of time that these parasites have been passaged in a laboratory setting, compared to other microsporidian species.

We also identified a *Nematocida* high-mobility group (HMG) domain gene, which is part of a candidate mating type locus previously identified in other microsporidia, based on comparisons to zygomycetes (Lee et al. 2008). In each of the three *Nematocida* genomes there was a single copy of this gene (NEPG\_01497/NEQG\_01453/NERG\_01188). There were no amino acid

changes in this gene between the two alleles in ERTm2, however there were differences between the gene in ERTm2 and in ERTm1/m3.

### **Phylogenomics of the Microsporidia and *Nematocida***

To better understand genome evolution in *Nematocida* and the Microsporidia overall, we compared *Nematocida* genomes to those of seven additional microsporidia, 13 other fungi, and an outgroup, the choanoflagellate *Monosiga brevicollis*. We then estimated a maximum likelihood phylogeny based on 53 single-copy core orthologs (Figure 3). Microsporidia were resolved as the earliest branching clade within the sequenced fungi. This result is consistent with a parallel phylogenetic analysis (Capella-Gutierrez et al. 2012) and with the prior report that microsporidia branch early within the fungal kingdom, and group with the most basally known fungus, *Rozella allomycis* (James et al. 2006). The three chytrid fungi (Chytridiomycota and Blastocladiomycota) were resolved in our phylogeny as the next most basal monophyletic clade. Based on these data, the hypothesis that Microsporidia could be a sister-group to Mucoromycotina including *Rhizopus*, as had been previously proposed based on gene order conservation (Lee et al. 2010), is rejected ( $p=7 \times 10^{-46}$ , AU test (Shimodaira 2002)). Placement of the Microsporidia as sister group to the chytrids was also rejected ( $p=5 \times 10^{-49}$ , AU test). Within the Microsporidia, *Nematocida* are the most basal clade in our analysis. Most relationships are consistent with previously published microsporidian phylogenies (Troemel et al. 2008), although *Antonospora* is positioned less closely to *Nematocida* and the position of *Vavraia* is more basal than previously estimated (Troemel et al. 2008). However, previous phylogenetic analysis used *V. oncoperae* for comparison, while *V. culicis* was used here. These two species may not form a monophyletic genus, which would explain the above discrepancies.

This phylogeny also highlights the large divergence within the Microsporidia, and between *Nematocida* and other species in particular. While *N. parisii* and *Nematocida* sp1 orthologs share 62% identity on average, *N. parisii* and *E. cuniculi* orthologs share only 31% identity (Figure S5), suggesting that only highly conserved genes will be detected as orthologs given such high average divergence. Mapping orthologs onto the phylogenetic tree identified 882 core microsporidian genes (ortholog clusters present in 8 or more genomes), with 137 of these being specific to Microsporidia (Figure 3). However, the majority of Microsporidia-specific genes lacked any functional assignment. We therefore altered our search to find ortholog clusters with a predicted function (i.e., PFAM domains or kinase annotations) that were either unique to Microsporidia or shared between Microsporidia and other basal groups, including the choanoflagellate *Monosiga brevicollis* and the basal fungi Chytridiomycota and *Rhizopus oryzae* (Table 2, Supp. Text).

Analysis of microsporidia-specific PFAM domains identified the nucleoside phosphate transporter (Npt) proteins, which have been implicated in nucleotide “stealing” from host cells (Table 2). The obligate intracellular bacterial pathogen *C. trachomatis* encodes two such transporters: Npt1 specifically transports ATP/ADP while Npt2 transports all four ribonucleotides (Tjaden et al. 1999). Four such transporters have previously been described in the microsporidian *E. cuniculi*, with three expressed on the plasma membrane to import ATP from the host, and one on the mitosomal membrane to import ATP for Fe/S clustering in this genomeless organelle (Goldberg et al. 2008; Tsaousis et al. 2008). The four *E. cuniculi* transporters have a preference for ATP, but may import other nucleotides as well. The

*Nematocida* genome only encodes two Npt transporters. Phylogenetic analysis indicates that a single ancestral transporter exhibited organism-specific duplications in *Nematocida*, *Antonospora*, and *Vavraia* (Figure S6). Further duplications led to greater numbers of these transporters in different microsporidia species, such as *E. cuniculi*. We speculate that all microsporidian genomes encode at least one transporter to import host nucleotides into the parasite to fuel parasite growth and DNA replication, and one additional transporter to import ATP from parasite cytoplasm into the mitosome. In *N. parisii* (ERTm1) both nucleotide transporters are expressed at high levels throughout infection (Table S5). One transporter (NEPG\_01813) was among the most highly expressed genes at 8 hpi, while the other transporter (NEPG\_01539) was expressed very highly later at 40 hpi during spore formation, and was also expressed in isolated spore RNA. Perhaps NEPG\_01813 imports nucleotides from the host cell into the parasite during early stages of infection for rapid DNA replication, while NEPG\_01539 imports ATP from the parasite cytoplasm into the mitosome during later stages of infection.

Our PFAM analysis also identified a putative nucleoside H<sup>+</sup> symporter, which is specific to Microsporidia (Table 2). Expression of this nucleoside transporter in *N. parisii* (ERTm1) (NEPG\_00699) can be detected throughout infection (Table S5). This symporter is homologous to the *E. coli* gene NupG, a broad-specificity transporter of purine and pyrimidine nucleosides that uses the proton motive force to drive nucleoside transport (Xie et al. 2004). Evolutionarily unrelated equilibrative nucleoside transporters are used by eukaryotic parasites such as *Trypanosoma* species that, like microsporidia, lack complete nucleoside synthesis pathways and need to import nucleosides from their hosts (Sanchez et al. 2002). However, unlike the equilibrative nucleoside transporter family, which is widely distributed across eukaryotes and is

absent from bacteria, the nucleoside H<sup>+</sup> symporter family is found primarily in bacteria, arthropods, and a small number of other eukaryotes (<http://pfam.sanger.ac.uk/family/PF03825>). The lack of any nucleoside H<sup>+</sup> symporter in basal fungi suggests that microsporidia acquired this transporter via horizontal gene transfer, although the exact source of the transfer is unclear.

Notably, we found a specific loss of the Retinoblastoma (RB) tumor suppressor gene in the Microsporidia phylum (Table 2, Figure 3). In most eukaryotes, the RB-E2F pathway regulates the cell cycle to promote cell proliferation and is composed of two activator proteins called E2F and DP, and a repressor protein called RB. The RB-E2F pathway is thought to have been present in the last common ancestor of eukaryotes, but has been lost in the Dikarya, having been replaced by an evolutionarily unrelated but functionally analogous cell cycle complex (Schaefer and Breeden 2004; Cao et al. 2010; Cross et al. 2011). In plants and animals, the E2F/DP activators are normally kept in check by the inhibitor RB, which is encoded by a tumor suppressor gene lost in most cancers (Polager and Ginsberg 2009). Microsporidia encode E2F and DP proteins, but lack recognizable RB proteins (through either orthology or translated BLAST against microsporidian genomes at a cutoff of 1e-5) indicating that microsporidia have lost the RB inhibitor.

In addition, we identified other domains and protein kinases shared between the microsporidia and basal fungi, as well as candidate *Nematocida* structural proteins including spore wall proteins and polar tube proteins (see Supp. Text and Table S10). We also identified a large number of genes that are specific to the *Nematocida* genus, as we could not find orthologs of these genes in other microsporidia or in other fungi (Figure 3). Only 11% of these *Nematocida*-

specific genes could be assigned PFAM domains, leaving most without functional assignment. In particular, we found a very large *Nematocida*-specific gene family (Figure S7A) that was not similar to any other genes in Genbank, but contained a conserved domain of 132 amino acids (Figure S7B). Using a Hidden Markov Model (HMM) profile built for this domain, we identified 107 members of this gene family in each of ERTm1 and ERTm3, and 160 members in ERTm2, representing 4-6% of the total predicted gene content of *Nematocida*. Proteins in this *Nematocida*-specific family have an average size of 263 aa, and often contain predicted secretion signals; in particular, 31 proteins in this family are predicted to be secreted in ERTm1, 26 in ERTm3, and 62 in ERTm2. Phylogenetic analysis showed a large number of species-specific expansions in this gene family. One large clade that is expanded within *Nematocida* is significantly enriched for genes predicted to be secreted (Figure S7A, lower half of tree) ( $p < 0.0001$ , chi squared test). Large families of secreted proteins are characteristic of virulence factors in other pathogens (Torto-Alalibo et al. 2010).

### ***Nematocida* metabolism**

Microsporidia are notable for their extremely reduced metabolic capability. Similar to analyses of other microsporidia (Keeling et al. 2010), we found that *Nematocida* genomes do not encode for components of an oxidative phosphorylation pathway or an intact tricarboxylic acid (TCA) cycle pathway, but do encode some core carbon pathways such as glycolysis (see Supp. Text and below). In addition, seven of eight key pentose phosphate pathway enzymes are present in *Nematocida* genomes, but a transaldolase ortholog appears to be absent, as in other microsporidian genomes. However, we found that *Nematocida* and other microsporidia encode for Class I aldolase (also known as aldolase B) (Table 2) and may use it for glycolysis, in

contrast to other fungi, which use Class II aldolase. Notably, the pentose phosphate pathway transaldolase is part of the Class I aldolase family within the aldolase superfamily (<http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.d.b.bb.html>) (Jia et al. 1996). Therefore, it is possible that microsporidia use Class I aldolase both in glycolysis and the pentose phosphate pathway, as this ‘double-duty’ strategy would be an efficient use of the microsporidian enzymatic repertoire.

*Nematocida* and other microsporidia have very limited biosynthetic and degradative capabilities, as compared to other fungi (see Supp. Text) but have retained a gene that encodes CTP synthase, which converts UTP into CTP, a nucleotide used for RNA/DNA synthesis and lipid synthesis (Chang and Carman 2008). CTP synthase is also present in the reduced genome of *C. trachomatis*, which encodes an active CTP synthase despite being able to import CTP (Wylie et al. 1996). Perhaps rapidly replicating intracellular pathogens need to supplement cytosine import, or balance nucleotide pools using CTP synthase. Alternatively, CTP synthase may also play a structural role, as it has been shown to form an intracellular filament structure conserved from bacteria to mammals (Ingerson-Mahar et al. 2010; Noree et al. 2010).

### **Secretion signal sequences in microsporidian hexokinases**

Hexokinase catalyzes the conversion of glucose to glucose-6-phosphate. Glucose-6-phosphate can then enter the glycolytic pathway, or enter the pentose phosphate pathway (Figure 4B). In our metabolic profiling we noticed that hexokinase was the only highly expressed glycolytic enzyme expressed at 8 hpi (Figure S8A, Table S5). This early expression suggests that hexokinase could perform a function distinct from other glycolytic enzymes. We found that

hexokinase genes in all three *Nematocida* genomes have a predicted secretion signal, while other enzymes in the glycolytic pathway do not (Table S5, Figure S8A). This observation suggests that *Nematocida* hexokinase could be secreted out of the parasite cell and into the host cell.

To investigate whether hexokinase secretion into the host cell may be a general pathogenic strategy of microsporidia, we examined whether secretion signals are present in other microsporidia hexokinases. Indeed, of six microsporidia genomes with annotated hexokinase genes, five contain predicted secretion signal sequences (Figure S8B). Protein alignment of these hexokinases indicates that conservation of a secretion signal is not simply due to amino acid conservation in the amino-terminus (Figure S8B), and thus may reflect functional conservation. We also analyzed hexokinases from the 12 other fungi in our analysis, and none of these enzymes had predicted secretion signal sequences. Thus, secreted hexokinase enzymes appear specific to microsporidia, and may be a conserved adaptation of these intracellular parasites to their host cells.

To experimentally determine whether these hexokinase secretion signal sequences could direct proteins through fungal secretory machinery, we tested these sequences in a yeast secretion trap system, which uses a simple growth assay to measure secretion in another fungus, *Saccharomyces cerevisiae* (Lee et al. 2006). Using this assay, the hexokinase secretion signal sequences from *N. parisii* (ERTm1 and ERTm3), as well as *Nematocida* sp1 (ERTm2) acted as functional secretion signal sequences (Figures 4A and S8C). Furthermore, we tested hexokinase secretion signal sequences from three other microsporidia species: *E. cuniculi*, *A. locustae*, and *N. ceranae*. In all cases we found that these secretion signals could traffic through the secretory

pathway of *S. cerevisiae*. The negative control *S. cerevisiae* amino-terminal domain from either hexokinase 1 (HXK1) or HXK2 did not direct secretion, nor did the amino-terminal sequence from an *N. parisii* (ERTm1) GTPase that is not predicted to be secreted, indicating specificity of this assay. In summary, our data suggest that during infection microsporidia hexokinases are secreted outside of the pathogen into the host cell cytoplasm, where they would be ideally situated to boost the production of host metabolites that could be used by these pathogens for replication.

## **Discussion:**

Our analyses provide insights into events that may have facilitated the emergence of Microsporidia, the largest phylum of obligate intracellular parasites. First, we find that microsporidia hexokinases have been modified to include a functional secretion signal, which likely exports it into host cells (Figure 4A and 4B). This finding could explain previous results with the microsporidian species *Nosema gryllii* where activity of several glycolytic enzymes was detected in isolated pathogen cells, but hexokinase activity was distinctively absent (Dolgikh 2000). Furthermore, it has been reported that microsporidia infection causes depletion of host glycogen and rapid uptake of glucose (Metenier and Vivares 2001), events that could be promoted by delivery of the parasite hexokinase into host cells. Hexokinase catalyzes the first step in both glycolysis and the pentose phosphate pathway. Therefore, microsporidia hexokinase activity within host cells could increase host synthesis of building blocks, such as nucleotides, amino acids and lipids, necessary for the rapid growth of these parasites (Figure 4B).

Our data show that during *N. parisii* infection, *C. elegans* intestinal cells become completely consumed by parasites, as visualized in Figure 1D. This figure highlights the conversion of host intestinal cells into parasite nucleotides, with parasite ribosomal RNA shown in red and parasite DNA in blue. Because *N. parisii* lacks almost all nucleotide biosynthetic capability (Supp. Text) there is considerable demand on host metabolism to provide these nucleotides. In order to meet these demands, microsporidia hexokinase would be ideally situated to redirect host cells towards anabolic metabolism (Figure 4B). This host metabolic state shares similarities with the Warburg effect, where cancer cells switch to glycolysis to help meet their proliferative demands for nucleotides, amino acids and lipids (Vander Heiden et al. 2009). Interestingly, the Warburg effect has been observed in host cells infected by Kaposi's sarcoma herpesvirus, another obligate intracellular pathogen (Delgado et al. 2010).

We speculate that an important step toward becoming a successful obligate intracellular pathogen is to increase the biosynthetic output of the host. For microsporidia, this step may have been the conversion of an amino-terminal hydrophobic domain found in hexokinase (John et al. 2011) into an amino-terminal hydrophobic secretion signal, which would intertwine the metabolisms of host and microbe. This marriage may have increased the fitness and shaped the pathogenesis of microsporidia, while similar events in other host/microbe relationships may have led to symbiosis, or even the evolution of organelles such as mitochondria (Valdivia and Heitman 2007; Keeling 2011).

Our phylogenomic analysis revealed that Microsporidia lost the RB tumor suppressor gene (Figure 3), which may explain the rapid intracellular growth of microsporidia within the host cell

(Figure 4B). RB is a key cell cycle inhibitor that stalls cells in the first gap phase (G1) of the cell cycle, thus delaying entry into synthesis (S) phase, when DNA replication occurs (Cao et al. 2010). In humans loss of RB leads to an accelerated cell cycle and a dramatically increased cancer risk. In fact, most cancer cells have lost the RB gene (Polager and Ginsberg 2009). Consistent with rapid entry into S phase, we find that *N. parisii* has an estimated 2.9 h to 3.3 h doubling time inside *C. elegans* host cells, quite similar to yeast growing in rich media. It is possible that microsporidia have adopted a cell cycle inhibitor unrelated in sequence to RB: indeed, microsporidia must halt the cell cycle at some point, in order to differentiate into spores. However, we speculate that loss of cell cycle inhibition early during infection may have allowed microsporidia to rapidly take advantage of imported nucleotides (see below) in order to outpace host cells. In addition, it is interesting to consider that the loss of a G1 cell cycle constraint in response to DNA damage would result in an increased mutation rate, perhaps explaining the longer branch lengths of microsporidia relative to other fungi in a phylogenetic tree (Figure 3).

Microsporidian genomes encode unique nucleoside and nucleotide transporters that could draw host-synthesized nucleosides into the parasite cell (Figure 4B). Our phylogenomics studies identified Npt nucleotide transporters, as previously shown to be present in the microsporidian *E. cuniculi*, as well as the obligate bacterium *C. trachomatis*. Furthermore, we identified a new class of nucleoside transporters in the microsporidia, which we propose were acquired by horizontal gene transfer (Table 2). Transporters like these, as well as other microsporidia-specific genes, are candidate drug targets for both medical and agricultural purposes.

We also identified a CTP synthetase as the only nucleotide synthesis gene retained in the highly reduced genomes of the microsporidia. This gene is also conserved in the reduced genome of the bacterium *C. trachomatis*, and these retentions suggest that it may perform a distinct function in addition to biosynthesis. CTP synthase has recently been proposed to constitute a cytoskeletal system, as it forms filaments in diverse organisms and regulates cell shape (Ingerson-Mahar et al. 2010; Noree et al. 2010). Maintenance of enzymes that serve multiple functions such as CTP synthase and perhaps Class I aldolase would be consistent with the overall economy of these highly reduced genomes. Notably, retention of CTP synthase, as well as acquisition of transporters, are features in common between the reduced prokaryotic genome of *C. trachomatis* and the reduced eukaryotic genomes of the microsporidia, suggesting convergent evolutionary strategies between these unrelated obligate intracellular pathogens. Furthermore, up-regulation of CTP synthase activity has been observed in tumors and leukemic cells (Williams et al. 1978; Kizaki et al. 1980; Weber et al. 1980), again highlighting the similarities between obligate intracellular pathogens and cancer cells.

Finally, microsporidia may generate genetic diversity by undergoing a rare sexual cycle and recombination, or by mitotic recombination resulting in loss of heterozygosity, as we found evidence to support this in *Nematocida*. Previous studies in other species of microsporidia have suggested the presence of a sexual cycle (Becnel et al. 2005). Notably, many species of microsporidia are dikaryotic (i.e. have two nuclei paired in one cell). The life cycle of *N. parisii* is less well-characterized than other species, but it involves meronts of at least two distinct types that contain between one to several nuclei within one cell (Troemel et al. 2008). These nuclei are often unpaired (monokaryotic), although two nuclei in close proximity are occasionally

observed, and might represent the paired nuclei of a dikaryotic stage. Perhaps these nuclei fuse as part of a conventional sexual cycle that undergoes reductive division by meiosis, or as part of a parasexual cycle that undergoes reductive division by chromosomal loss, as observed in the fungal pathogen *Candida albicans* (Bennett and Johnson 2003). Our data provide the first molecular support for mating and recombination occurring in microsporidia, and we speculate that it may allow them to repair mutations, or alternatively to propagate mutations and increase virulence, similar to other fungal pathogens (Ni et al. 2011). In summary, we propose that rapid clonal proliferation using the cancer-like strategies described above, together with a rare sexual cycle, may complement each other to allow for rapid intracellular growth and successful spread of microsporidia throughout the animal kingdom.

## **Methods:**

### *Nematocida* genomic DNA isolation

All three *Nematocida* strains (*N. parisii* strains ERTm1 and ERTm3, and *Nematocida* sp1 strain ERTm2) were transferred into the *C. elegans* wild-type N2 strain for propagation and harvest (Troemel et al. 2008). See Supp. Methods for details on harvest and DNA extraction.

### *Nematocida* RNA isolation and RNA-seq library construction

*C. elegans* growth and infections were performed at 25°C and *N. parisii* spores were prepared as previously described (Estes et al. 2011). The *C. elegans* temperature-sensitive sterile strain *fer-15(b26);fem-1(hc17)* was used to prevent internal hatching of progeny during the progression of infection. Synchronized *fer-15;fem-1* L1s were added to 10-cm NGM plates seeded with a lawn of *E. coli* OP50 bacteria, grown at the restrictive temperature of 25C for 24 hours, and then  $3 \times 10^6$  *N. parisii* ERTm1 spores diluted in M9 media (or M9 media alone for uninfected controls)

were spread over the entire surface of these plates. Infected and control animals were harvested at the appropriate times and total RNA was extracted using TRI Reagent (Molecular Research Center, Inc). See Supp. Methods for details on assessing life cycle and infection kinetics using microscopy, RNA quality analysis, and RNA isolation from spores. Strand-specific libraries were constructed for all RNA samples using the dUTP second strand marking method (Parkhomchuk et al. 2009; Levin et al. 2010). For detailed protocols, see Supp. Methods.

#### Genome sequencing, assembly, and single nucleotide polymorphisms (SNPs)

DNA for each genome was sequenced using 454 FLX-Titanium technology for two whole genome shotgun libraries, 400 base fragment and 2.7 kb paired-end reads. Reads were screened by Blast against the non-redundant Genbank nucleotide (nt) database to identify possible contamination; of 1,000 reads examined for each species, <2% matched the *E. coli* genome and < 0.1% matched the *C. elegans* genome. 454 reads were assembled with Newbler (MapAsmResearch-03/15/2010). A very high fraction (>95%) of reads were assembled. Contamination with *C. elegans* and *E. coli* was removed using Newbler's -vs option. Assemblies were deposited in Genbank for *N. parisii* ERTm1 (AEFF01000000), *N. parisii* ERTm3 (AE0001000000), and *Nematocida* sp1 (AERB01000000). For further details, see Supp. Methods.

Illumina sequence for two ERTm2 samples was used to identify polymorphisms using GATK. We generated 101 bp reads for the same sample used to generate the genome sequence using 454 data, and for a second sample serially passaged for 14 weeks (as described above). For further details see Supp. Methods.

### Gene prediction and RNA-seq analysis

For ERTm1 and ERTm2, an initial set of protein coding genes was predicted using Prodigal (Hyatt et al. 2010). For ERTm3, the initial gene set was generated by a synteny-based transfer of the gene set from ERTm1; transferred genes with internal frameshifts were replaced by Prodigal predictions for that locus. RNA-seq was used to refine the gene predictions for ERTm1 (Supp. Methods). To calculate the coverage of all RNA-seq data (including spores) after we had ascertained there were no spliced genes, we realigned all sequences to the ERTm1 genome using BWA (Li and Durbin 2009), and then calculated fragments per kilobase of transcript per million fragments mapped (FPKM) (Trapnell et al. 2010). For further details see Supp. Methods.

### Gene clustering and evolutionary analysis

Orthologous gene families were identified for the three *Nematocida*, seven other microsporidia, 12 fungi and one choanoflagellate outgroup (see Supp. Methods) using ORTHOMCL version 1.4 with a Markov inflation index of 1.5 and a maximum e-value of  $1 \times 10^{-5}$ . To estimate a phylogeny, we selected 53 orthologs present as single copy in all genomes. Amino acid sequences of these clusters were aligned using MUSCLE (Edgar 2004) and poorly aligned regions were trimmed using trimAl under default settings (Capella-Gutierrez et al. 2009). We then estimated a phylogeny using the PROTGAMMABLOSUM62 model in RAxML (Stamatakis 2006) with 1,000 bootstrap replicates. We compared this multi-ortholog-based phylogeny to phylogenies with Microsporidia repositioned to be 1) sister-group to *Rhizopus*, and 2) sister-group to the chytrids, with the approximately unbiased test (Shimodaira 2002) as implemented in CONSEL

(Shimodaira and Hasegawa 2001), using the single copy core amino acid alignments used to generate the organismal phylogeny. For further details see Supp. Methods.

### Yeast secretion trap analysis of candidate signal sequences

To test whether predicted signal sequences would direct trafficking through the *S. cerevisiae* secretory pathway we used a yeast secretion trap system as described before (Lee et al. 2006). Briefly, we PCR amplified the first 85 amino acids from candidate genes and fused them at their C-terminus to the *S. cerevisiae* invertase reporter gene using vector pYST-1 (a kind gift from Jocelyn Rose). The DNA templates for PCR amplification were extracted from spores (*N. ceranae* spores were a kind gift of James Nieh and Guntima Suwannapong, *E. cuniculi* spores were a kind gift of Quanshun Zhang and Saul Tzipori and *A. locustae* spores were purchased from M&R Durango). The resulting fusions were transformed into an invertase deletion strain of *S. cerevisiae* SEY6210 (ATCC96099) and the transformants plated onto synthetic leucine drop-out medium with 2 µg/mL antimycin A (Sigma A8674) and either glucose or sucrose as the sole carbon source. Secretion of the invertase fusion protein directed by the N-terminal candidate gene sequence rescues the deletion mutant and results in yeast growth on sucrose.

### **Data access:**

Assemblies and annotations were submitted to GenBank under the following accession numbers: *N. parisii* ERTm1 (AEFF01000000), *N. parisii* ERTm3 (AE0001000000), *N. sp1* ERTm2 (AERB01000000). SNPs for ERTm2 were submitted to dbSNP and scheduled for Build 138 release (handle: BROAD-GENOME BIO, sbid 1057048, ss# 521825843-521933378 (not inclusive)). SNPs for ERTm1 and ERTm2 were submitted to dbSNP and release scheduling is pending (handle: BROAD-GENOME BIO).

## Acknowledgments:

We acknowledge the Broad Institute Sequencing Platform for generating all DNA and RNA sequence described here. We thank Brian Haas for assistance with defining UTRs from the RNA-sequence, Jeremy Zucker for metabolic pathway prediction, Hilary Morrison for providing the *A. locustae* genome sequence prior to publication, Sharadha Sakthikumar for assistance with Figure 2 preparation, James Nieh and Guntima Suwannapong for *N. ceranae* spores, Quanshun Zhang and Saul Tzipori for *E. cuniculi* spores, and Nicholas Rhind, Robert Luallen, Tiffany Dunbar, Ethan Bier, Suzannah Szumowski, Diego Martinez, Zehua Chen, and Christina Hull for comments on the manuscript. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C and funding to Tulane National Primate Research Center via OD011104. This work was supported by the Irvington Institute Fellowship Program of the Cancer Research Institute postdoctoral fellowship to M.A.B., and NIAID R01 AI087528, Center for AIDS Research Developmental Grant, the Searle Scholars Program, Ray Thomas Edwards Foundation, David & Lucile Packard Foundation fellowship to E.R.T.

## Figure legends:

### Fig. 1. Characterization of *N. parisii* infection stages in *C. elegans*.

(A) Diagram of infection stages. *N. parisii* infects *C. elegans* intestinal cells (green) where it grows intracellularly from small mononucleate sporoplasms to large, multinucleate meronts. Meronts develop into spores, which at late stages of infection can be enclosed within vesicles.

(B) Representative DIC images of animals are shown displaying distinct infection-associated

phenotypes with boxed areas enlarged and shown as insets. Small spore (arrow), large spore (arrowhead). **(C)** *N. parisii*-specific FISH (red) and DAPI (blue) staining of animals are shown at different stages of infection. Note: although animals do not exhibit infection symptoms by DIC at 8 hpi, sporoplasm is visible by fluorescent in situ hybridization (FISH) at 8 hpi. For B and C, scale bar = 10  $\mu$ m. **(D)** The Y-axis on left indicates the fraction of animals in a population exhibiting specific symptoms of infection visualized by DIC ( $n > 22$  animals assayed per timepoint) (see Supp. Methods). RNA sample collection times are indicated on the graph: 1) 8 hpi, during the sporoplasm stage as observed by FISH, 2) 16 hpi, during the ‘early meront’ stage, 3) 30 hpi, during the ‘late meront’ stage, 4) 40 hpi, when spores have just started forming, and 5) 64 hpi, when *Nematocida* spores can be found within membrane-bound vesicles. The Y-axis on the right indicates the contribution of *N. parisii* reads to total reads from RNA-seq analysis.

**Fig. 2. Distribution of SNPs in *Nematocida* highlights loss of heterozygosity.**

**(A)** The frequency of SNPs for sliding windows of 5 kb is plotted across the 11 largest scaffolds of *Nematocida* sp1 (ERTm2). Values on the x-axis correspond to scaffold position in kb; values on the y-axis correspond to the SNP frequency (SNPs/kb). The pattern of the initial isolate (red) is nearly identical to the passaged isolate (blue). Large regions exhibiting loss of heterozygosity are present in terminal regions of scaffolds 1, 2, and 4. **(B)** The frequency of SNPs for sliding windows of 5kb is plotted across the 11 largest scaffolds of *N. parisii* (ERTm1). Values on the x-axis correspond to scaffold position in kb; values on the y-axis correspond to the SNP frequency (SNPs/kb). The pattern of SNPs for ERTm1 (blue) and ERTm3 (red) are highly similar.

**Fig. 3. Phylogeny and ortholog content of Microsporidia and Fungi.**

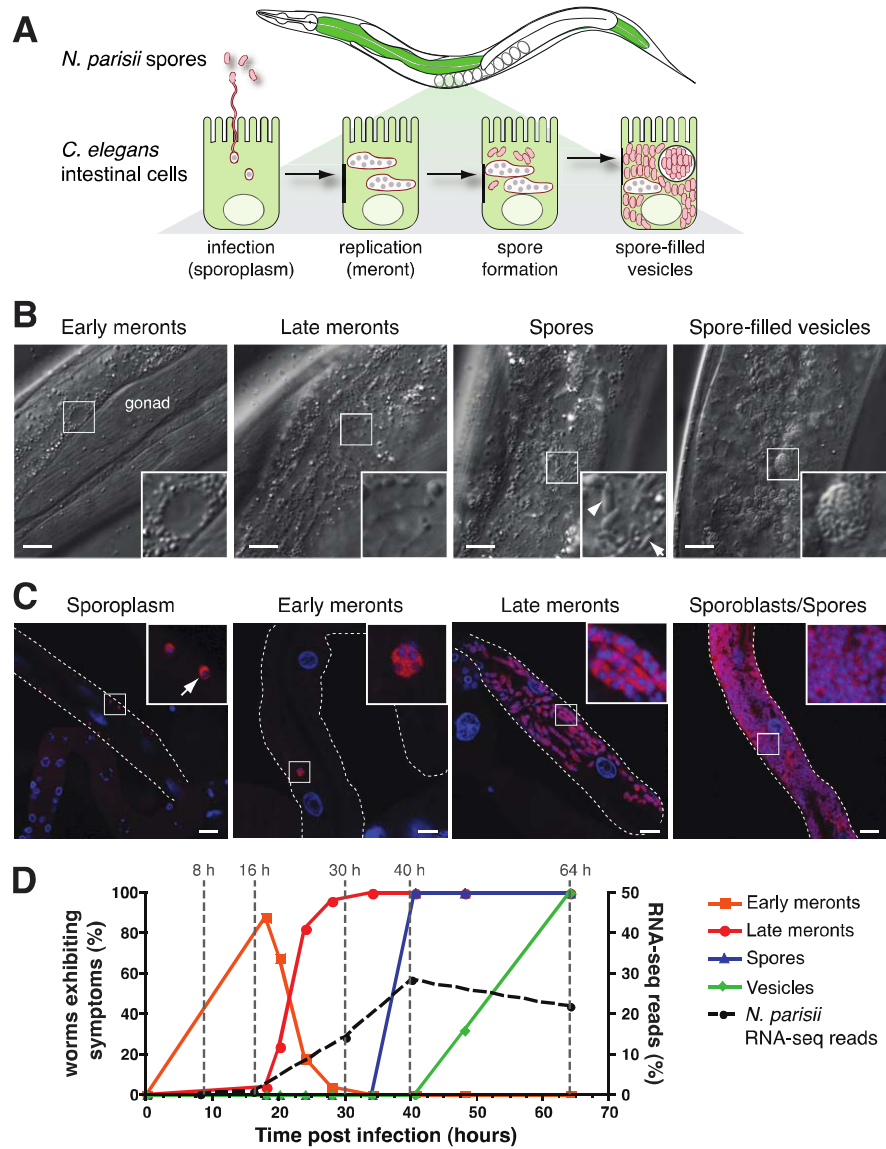
The phylogeny was inferred by maximum likelihood using RAxML (Stamatakis 2006) based on the concatenated amino acid sequences of 53 single copy orthologs shared by all taxa. Bootstrap values above 50% are listed above respective nodes. Ortholog counts are shown as bar graphs to the right of the tree and are divided into six categories: core (green, found in all genomes), shared (blue, found in at least two genomes excluding other Microsporidia and *Nematocida* categories that follow), Microsporidia-specific (orange), *Nematocida*-specific (red), *N. parisii*-specific (purple), and unique (yellow). Presence of E2F, DP, and RB orthologs is indicated in table on the right, with E2F and DP activators in green, and the RB inhibitor in red. Dikarya cell cycle regulation circuitry is evolutionarily distinct from the RB-E2F pathway (Cao et al. 2010).

**Fig. 4. Microsporidia hexokinase secretion signals can direct protein trafficking through the yeast secretion system.**

(A) A yeast secretion trap system was used to test the ability of microsporidia hexokinase secretion signals to direct secretion in a fungal system. Ten-fold serial dilutions of *S. cerevisiae* strains containing candidate signal sequences were grown either on media containing sucrose (left) or glucose (right). Growth on sucrose indicates a functional signal sequence (boxed in red).

(B) Model for evolutionary events that enable rapid intracellular growth of microsporidia: 1) hexokinase acquired a secretion signal sequence, likely directing this enzyme into the host cell where it can increase production of amino acids, lipids and nucleotides; 2) transporters were acquired, which can import host-synthesized nucleosides and nucleotides into the pathogen meront; and 3) RB was lost, which we speculate could lead to a rapid, but mistake-prone, cell cycle, with concomitant DNA and RNA synthesis supported by host derived nucleotides.

**Figures:**



**Fig. 1 Cuomo et al., 2012**

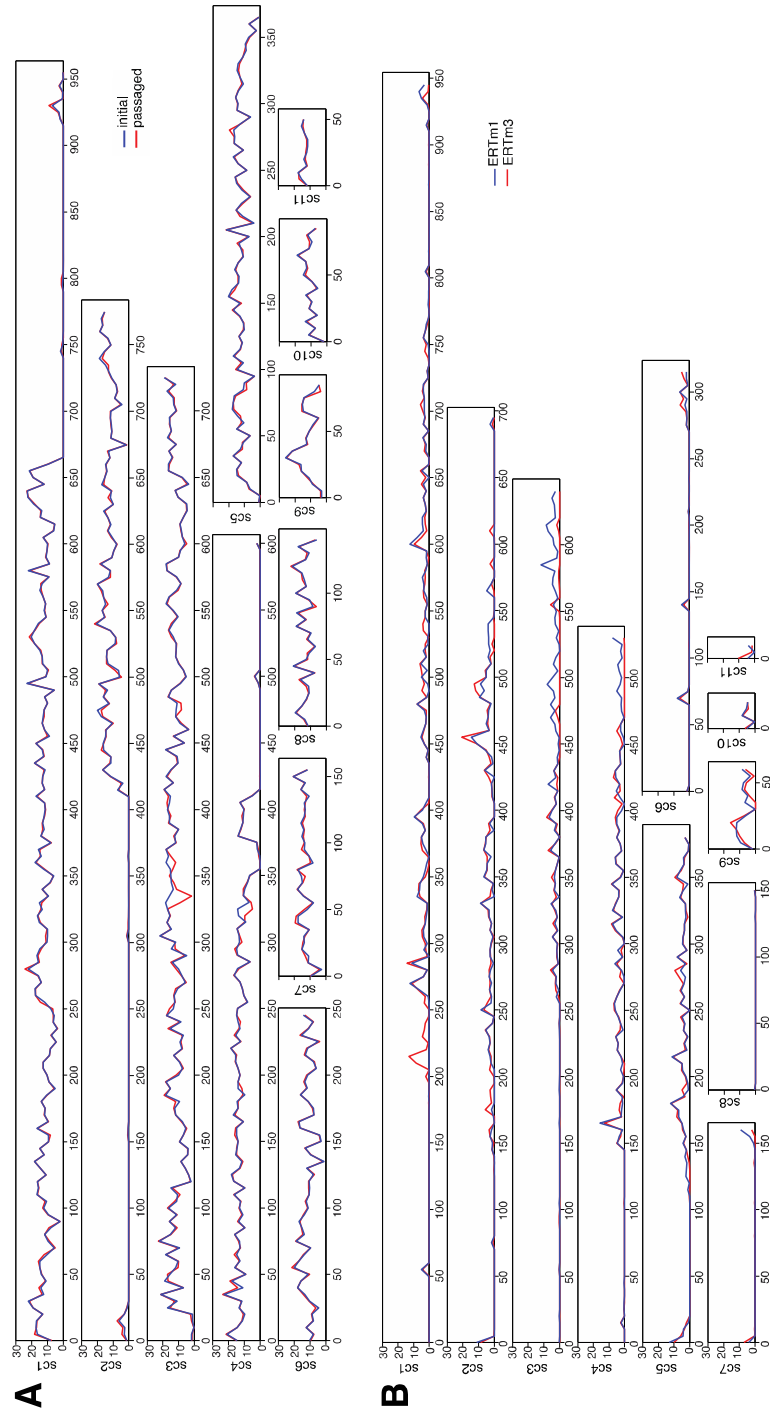


Fig. 2 Cuomo et al., 2012

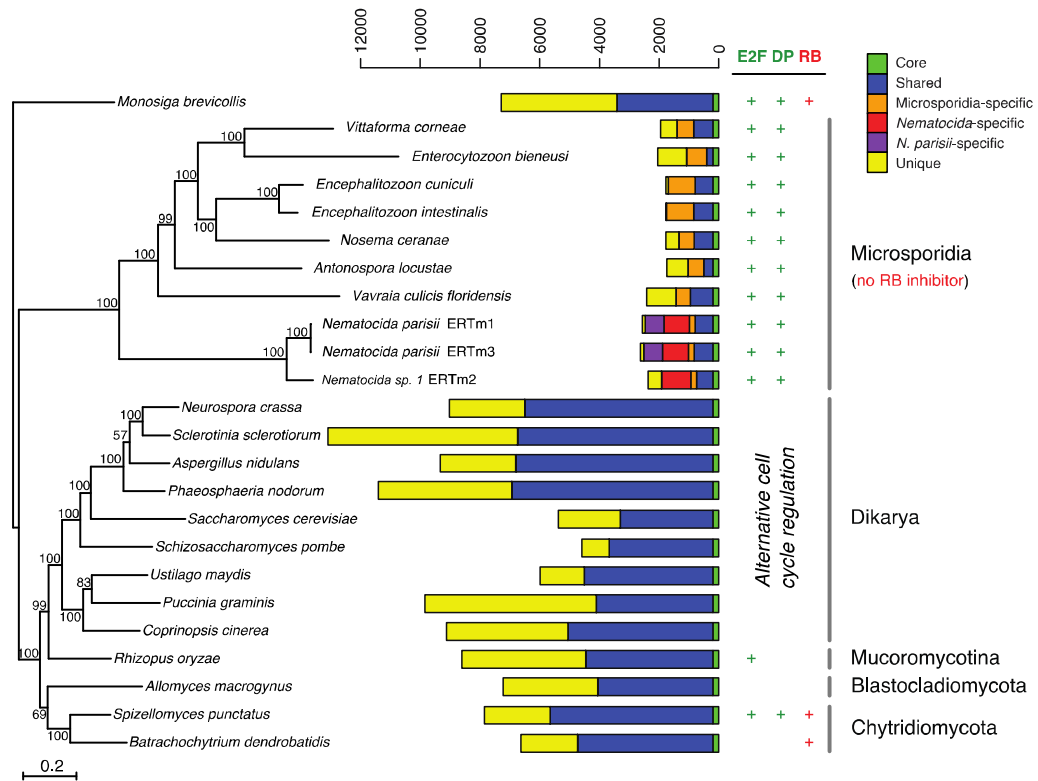


Fig. 3 Cuomo *et al.*, 2012

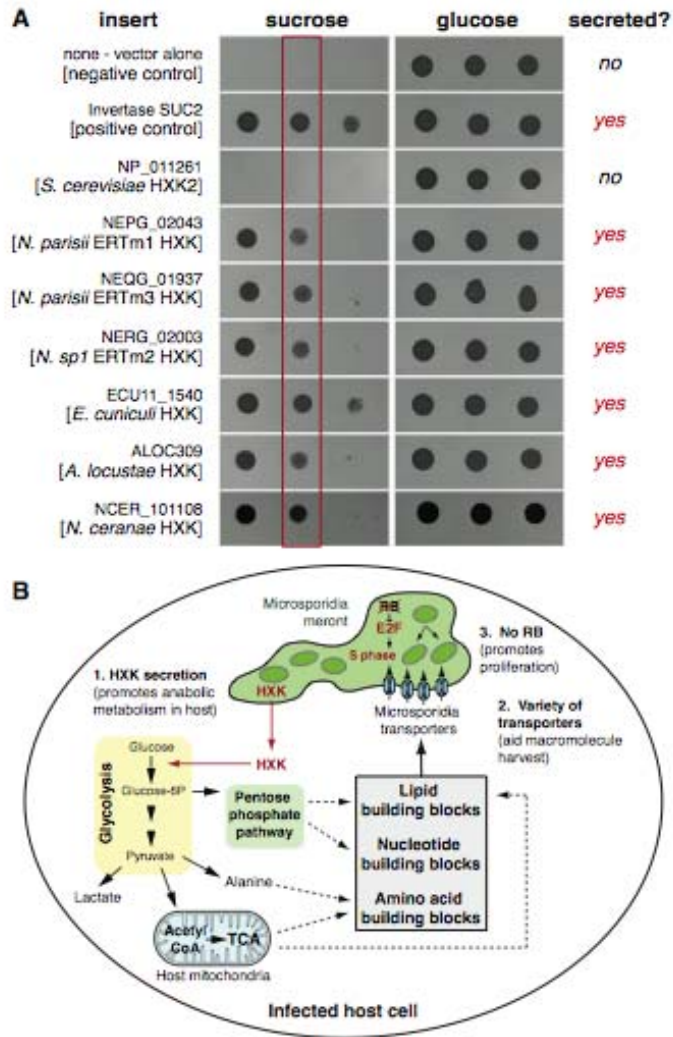


Fig. 4 Cuomo *et al.*, 2012

**Tables:****Table 1. Statistics of *Nematocida* genomes.**

	<i>N. parisii</i> ERTm1	<i>N. parisii</i> ERTm3	<i>N. sp1</i> ERTm2
Genome assembly (Mb)	4,075,448	4,147,597	4,700,711
Repeat content (%)	8.4%	8.2%	17.7%
GC content (%GC)	34.4%	34.5%	38.3%
Protein coding genes	2,661	2,726	2,770
Mean coding length (bp)	1,104	1,055	1,084
Mean 5'UTR (bp)	6		
Mean 3'UTR (bp)	52		
Mean non-coding length (bp)	418	457	569
Genes with RNA-seq evidence (#)	2,546		
Genes with signal peptides (#)	198	201	220
Genes with transmembrane domains (#)	710	702	680
Genes with pfam assignments (#)	1,104	1,101	1,142
Genes with GO assignments (#)	590	600	611

**Table 2. PFAM domains and protein kinases unique to Microsporidia with respect to other Fungi, or shared between Microsporidia and the basal Fungi.**

# of Genomes with Domain							
Microsporidia (10) <sup>^</sup>	Chytridiomycota (3)	Mucoromycotina (1)	Dikarya (9)	<i>Monosiga</i> (1)	PFAM Domain / Kinase Classification	Function	
10	0	0	0	0	PF03219 Nucleoside phosphate transporter (TLC ATP/ADP transporter)	Nucleotide transport	
10	0	0	0	0	PF00182 Chitinase class I	Chitin-related	
8	3	0	0	0	PF00274 Fructose-bisphosphate aldolase class-I	Glycolysis	
9	0	1	0	0	PF02121 Phosphatidylinositol transfer protein	Phospholipid trafficking and signaling	
10	3	1	0	0	PF01167 Tub family	Phospholipid trafficking and signaling	
9	1 <sup>a</sup>	0	1 <sup>a</sup>	0	PF03825 Nucleoside H <sup>+</sup> symporter	Nucleoside transport	
9	2	1	0	0	PF02224 Cytidylate kinase	Nucleotide metabolism	
10 <sup>b</sup>	1	0	0	1	PF08781 Transcription factor DP	RB-E2F cell cycle (DP)	
10	1	1	0	1	PF02319 E2F/DP family winged-helix DNA-binding domain	RB-E2F cell cycle (E2F)	
10	1	0	1 <sup>c</sup>	1	PF01139 Uncharacterized protein family 27		
9	3	1	0	1	WNK With-no-lysine kinase	Ion transport regulation, osmotic regulation	
9	3	1	1 <sup>d</sup>	1	DMPK Dystrophia myotonica-protein kinase	Ion transport regulation, cytoskeletal regulation	

<sup>^</sup> number of genomes analyzed in each column is listed in parentheses.

<sup>a</sup> present in non-orthologous copies in *Spizellomyces punctatus* and *Sclerotinia sclerotiorum*

<sup>b</sup> present as full-length RNA-seq based transcript in ERTm1 but missing from assembly

<sup>c</sup> present in *Phaeosphaeria nodorum*

<sup>d</sup> present in a non-orthologous copy in *Coprinopsis cinerea*

## References:

- Akiyoshi DE, Morrison HG, Lei S, Feng X, Zhang Q, Corradi N, Mayanja H, Tumwine JK, Keeling PJ, Weiss LM et al. 2009. Genomic survey of the non-cultivable opportunistic human pathogen, *Enterocytozoon bieneusi*. *PLoS pathogens* **5**(1): e1000261.
- Becnel JJ, White SE, Shapiro AM. 2005. Review of microsporidia-mosquito relationships: from the simple to the complex. *Folia Parasitol (Praha)* **52**(1-2): 41-50.
- Bennett RJ, Johnson AD. 2003. Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *Embo J* **22**(10): 2505-2515.
- Cao L, Peng B, Yao L, Zhang X, Sun K, Yang X, Yu L. 2010. The ancient function of RB-E2F pathway: insights from its evolutionary history. *Biol Direct* **5**: 55.
- Capella-Gutierrez S, Marcet-Houben M, Gabaldon T. 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol* **10**(1): 47.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**(15): 1972-1973.
- Chang YF, Carman GM. 2008. CTP synthetase and its role in phospholipid synthesis in the yeast *Saccharomyces cerevisiae*. *Prog Lipid Res* **47**(5): 333-339.
- Cornman RS, Chen YP, Schatz MC, Street C, Zhao Y, Desany B, Egholm M, Hutchison S, Pettis JS, Lipkin WI et al. 2009. Genomic analyses of the microsporidian *Nosema ceranae*, an emergent pathogen of honey bees. *PLoS pathogens* **5**(6): e1000466.
- Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nat Commun* **1**: 77.
- Cross FR, Buchler NE, Skotheim JM. 2011. Evolution of networks and sequences in eukaryotic cell cycle control. *Philos Trans R Soc Lond B Biol Sci* **366**(1584): 3532-3544.
- Delgado T, Carroll PA, Punjabi AS, Margineantu D, Hockenbery DM, Lagunoff M. 2010. Induction of the Warburg effect by Kaposi's sarcoma herpesvirus is required for the maintenance of latently infected endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **107**(23): 10696-10701.
- Didier ES, Weiss LM. 2011. Microsporidiosis: not just in AIDS patients. *Curr Opin Infect Dis* **24**(5): 490-495.
- Dolgikh VV. 2000. Activities of enzymes of carbohydrate and energy metabolism of the intracellular stages of the microsporidian, *Nosema grylli*. *Protistology* **1**(3): 87-91.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5): 1792-1797.
- Edlind TD, Li J, Visvesvara GS, Vodkin MH, McLaughlin GL, Katiyar SK. 1996. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. *Mol Phylogenet Evol* **5**(2): 359-367.
- Estes KA, Szumowski SC, Troemel ER. 2011. Non-lytic, actin-based exit of intracellular parasites from *C. elegans* intestinal cells. *PLoS pathogens* **7**(9): e1002227.

- Fantes PA. 1977. Control of cell size and cycle time in *Schizosaccharomyces pombe*. *Journal of cell science* **24**: 51-67.
- Gill EE, Lee RC, Corradi N, Grisdale CJ, Limpricht VO, Keeling PJ, Fast NM. 2010. Splicing and transcription differ between spore and intracellular life stages in the parasitic microsporidia. *Mol Biol Evol* **27**(7): 1579-1584.
- Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R, Embley TM. 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. *Nature* **452**(7187): 624-628.
- Hirt RP, Logsdon JM, Jr., Healy B, Dorey MW, Doolittle WF, Embley TM. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences of the United States of America* **96**(2): 580-585.
- Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.
- Ingerson-Mahar M, Briegel A, Werner JN, Jensen GJ, Gitai Z. 2010. The metabolic enzyme CTP synthase forms cytoskeletal filaments. *Nat Cell Biol* **12**(8): 739-746.
- James AB, Simpson TY, Chamberlain WA. 2008. Chlamydia prevalence among college students: reproductive and public health implications. *Sex Transm Dis* **35**(6): 529-532.
- James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J et al. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**(7113): 818-822.
- Jia J, Huang W, Schorken U, Sahn H, Sprenger GA, Lindqvist Y, Schneider G. 1996. Crystal structure of transaldolase B from *Escherichia coli* suggests a circular permutation of the alpha/beta barrel within the class I aldolase family. *Structure* **4**(6): 715-724.
- John S, Weiss JN, Ribalet B. 2011. Subcellular localization of hexokinases I and II directs the metabolic fate of glucose. *PloS one* **6**(3): e17674.
- Johnson RM, Evans JD, Robinson GE, Berenbaum MR. 2009. Changes in transcript abundance relating to colony collapse disorder in honey bees (*Apis mellifera*). *Proceedings of the National Academy of Sciences of the United States of America* **106**(35): 14790-14795.
- Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**(6862): 450-453.
- Keeling PJ. 2011. Endosymbiosis: bacteria sharing the load. *Current biology : CB* **21**(16): R623-624.
- Keeling PJ, Corradi N. 2011. Shrink it or lose it: balancing loss of function with shrinking genomes in the microsporidia. *Virulence* **2**(1): 67-70.
- Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S. 2010. The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. *Genome Biol Evol* **2**: 304-309.
- Keeling PJ, Fast NM. 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annual review of microbiology* **56**: 93-116.
- Keeling PJ, Luker MA, Palmer JD. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol Biol Evol* **17**(1): 23-31.
- Kizaki H, Williams JC, Morris HP, Weber G. 1980. Increased cytidine 5'-triphosphate synthetase activity in rat and human tumors. *Cancer Res* **40**(11): 3921-3927.

- Lee SC, Corradi N, Byrnes EJ, 3rd, Torres-Martinez S, Dietrich FS, Keeling PJ, Heitman J. 2008. Microsporidia evolved from ancestral sexual fungi. *Curr Biol* **18**(21): 1675-1679.
- Lee SC, Corradi N, Doan S, Dietrich FS, Keeling PJ, Heitman J. 2010. Evolution of the sex-related locus and genomic features shared in microsporidia and fungi. *PloS one* **5**(5): e10539.
- Lee SJ, Kim BD, Rose JK. 2006. Identification of eukaryotic secreted and cell surface proteins using the yeast secretion trap screen. *Nat Protoc* **1**(5): 2439-2447.
- Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**(9): 709-715.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
- Malik SB, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM, Jr. 2008. An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PloS one* **3**(8): e2879.
- Metenier G, Vivares CP. 2001. Molecular characteristics and physiology of microsporidia. *Microbes Infect* **3**(5): 407-415.
- Ni M, Feretzaki M, Sun S, Wang X, Heitman J. 2011. Sex in fungi. *Annu Rev Genet* **45**: 405-430.
- Noree C, Sato BK, Broyer RM, Wilhelm JE. 2010. Identification of novel filament-forming proteins in *Saccharomyces cerevisiae* and *Drosophila melanogaster*. *The Journal of cell biology* **190**(4): 541-551.
- Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**(18): e123.
- Polager S, Ginsberg D. 2009. p53 and E2f: partners in life and death. *Nature reviews Cancer* **9**(10): 738-748.
- Sanchez MA, Tryon R, Green J, Boor I, Landfear SM. 2002. Six related nucleoside/nucleobase transporters from *Trypanosoma brucei* exhibit distinct biochemical functions. *The Journal of biological chemistry* **277**(24): 21499-21504.
- Schaefer JB, Breeden LL. 2004. RB from a bud's eye view. *Cell* **117**(7): 849-850.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**(3): 492-508.
- Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**(12): 1246-1247.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21): 2688-2690.
- Stiller JW, Hall BD. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol* **16**(9): 1270-1279.
- Tjaden J, Winkler HH, Schwoppe C, Van Der Laan M, Mohlmann T, Neuhaus HE. 1999. Two nucleotide transport proteins in *Chlamydia trachomatis*, one for net nucleoside triphosphate uptake and the other for transport of energy. *J Bacteriol* **181**(4): 1196-1202.
- Torto-Alalibo T, Collmer CW, Gwinn-Giglio M, Lindeberg M, Meng S, Chibucos MC, Tseng TT, Lomax J, Biehl B, Ireland A et al. 2010. Unifying themes in microbial associations with animal and plant hosts described using the gene ontology. *Microbiology and molecular biology reviews : MMBR* **74**(4): 479-503.

- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.
- Troemel ER. 2011. New Models of Microsporidiosis: Infections in Zebrafish, *C. elegans*, and Honey Bee. *PLoS pathogens* **7**(2): e1001243.
- Troemel ER, Felix MA, Whiteman NK, Barriere A, Ausubel FM. 2008. Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *PLoS biology* **6**(12): 2736-2752.
- Tsaousis AD, Kunji ER, Goldberg AV, Lucocq JM, Hirt RP, Embley TM. 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. *Nature* **453**(7194): 553-556.
- Valdivia RH, Heitman J. 2007. Endosymbiosis: the evil within. *Current biology : CB* **17**(11): R408-410.
- Vander Heiden MG, Cantley LC, Thompson CB. 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, NY)* **324**(5930): 1029-1033.
- Weber G, Lui MS, Takeda E, Denton JE. 1980. Enzymology of human colon tumors. *Life Sci* **27**(9): 793-799.
- Williams BA. 2009. Unique physiology of host-parasite interactions in microsporidia infections. *Cellular microbiology* **11**(11): 1551-1560.
- Williams JC, Kizaki H, Weber G, Morris HP. 1978. Increased CTP synthetase activity in cancer cells. *Nature* **271**(5640): 71-73.
- Wylie JL, Berry JD, McClarty G. 1996. *Chlamydia trachomatis* CTP synthetase: molecular characterization and developmental regulation of expression. *Mol Microbiol* **22**(4): 631-642.
- Xie H, Patching SG, Gallagher MP, Litherland GJ, Brough AR, Venter H, Yao SY, Ng AM, Young JD, Herbert RB et al. 2004. Purification and properties of the *Escherichia coli* nucleoside transporter NupG, a paradigm for a major facilitator transporter sub-family. *Mol Membr Biol* **21**(5): 323-336.