Comparative genomics of the phylum Microsporidia: defining shared and specific genes in a phylum of emerging human pathogens

A white paper for Microbial Genome Sequencing submitted by Emily Troemel\textsuperscript{1} and Christina Cuomo\textsuperscript{2} on behalf of the Microsporidian Genomes Consortium.
\textsuperscript{1}University of California, San Diego, La Jolla, California \textsuperscript{2}Broad Institute, Cambridge, MA
1. Summary
Microsporidia comprise an exceptionally diverse class of over 1,200 species of intracellular pathogens that infect almost all animal phyla [1, 2]. At least 14 species of microsporidia can infect humans, and these infections can lead to death in immunocompromised patients. Vertebrate microsporidia most commonly infect the intestine, where they can cause severe, persistent diarrhea. However, they have been isolated from virtually all organ systems. The route of infection in most cases is oral, but microsporidia can also inoculate the eye and the skin, likely via wounds. *Enterocytozoon bieneusi* is the most common cause of microsporidian infections in humans, and there is a need for effective drugs to treat infections caused by this pathogen. This problem, together with the fact that microsporidia have been found in water sources, led the NIAID to add microsporidia to its list of Category B priority pathogens.

Microsporidia were originally classified as protozoa and were proposed to be ancient eukaryotes, largely due to their apparent lack of canonical eukaryotic features (such as typical mitochondria). Early phylogenetic analyses placed them deep within the eukaryotic radiation [3]. However, the application of more accurate phylogenetic methods has shown that these pathogens belong within the fungal kingdom and are highly specialized, reduced fungi [4, 5]. Recent studies indicate that these fungi are derived from zygomycetes and thus may retain a sexual cycle [6]. The success of this class of pathogens is evident by their remarkably wide host range and global distribution.

This white paper aims to provide genomic resources for studying microsporidia, which are one of the most ubiquitous but poorly understood classes of pathogens. Because of the poor clinical diagnostics available for microsporidia, the full extent of their effect on the US population has not been determined. Their prevalence is almost certainly under-reported because of these shortcomings. Our specific goals are to:

- identify genes that will provide better tools for diagnostics and therapeutics against microsporidia
- define a set of genes common to all microsporidia, including those genes specific to microsporidia
- perform comparative analysis of species with different hosts in order to identify genes that define host range
- determine subspecies strain variability in order to investigate a possible sexual cycle

The small genome size of these pathogens (many are half the size of *E. coli*) makes it feasible to analyze several genomes at relatively low cost. We have chosen the species below because it is possible to obtain significant quantities of pure sample, they are clinically relevant and/or provide good models for studying microsporidian pathogenesis. Specifically, we propose to sequence (see Table 1):

- Human strains: *Encephalitozoon cuniculi, Encephalitozoon intestinalis, Encephalitozoon hellem, Enterocytozoon bieneusi, Annecalla algerae, Vittaforma corneae*
- Model system strains: *Edhazardia aedis, Nucleospora salmonis, Pseudoloma neurophilia, Vavraia culicis, Nematocida parisi*

Analysis of several species in this phylum will provide insights into the core set of genes that are responsible for their specialized intracellular lifestyle and development. By comparing and
contrasting different strains we will also determine what is responsible for their diverse host range and developmental cycles. This project will provide the microsporidia community with genomic tools that will greatly facilitate research into these under-studied, but medically relevant group of organisms.

2. Medical Significance of microsporidia
The threat from microsporidia was underappreciated until the AIDS epidemic, when it was discovered in the 1980’s that HIV+ patients were afflicted with severe diarrhea due to microsporidian infections. Since that time, microsporidia have also caused death in transplant patients, and have been isolated from immunocompetent patients as well. Infections most commonly occur in the gastrointestinal tract, but also occur with some frequency in other organs, such as the lung, muscle, nervous tissue and eye.

*Encephalitozoon species*
*Encephalitozoon cuniculi* is among the most common species of microsporidia that infects humans, as well as other mammals. *E. cuniculi* infects intestinal epithelial cells, where it can cause serious diarrhea in immunocompromised patients. However, *E. cuniculi* can also disseminate to virtually all other organ systems, including brain, lung, kidney, biliary tract, eye and nervous system. For example, *E. cuniculi* was fairly recently identified in a fatal pulmonary infection in a bone marrow transplant patient [7]. Very recently, serious endocarditis in an immunocompentent patient was determined to be due to a large vegetative mass of *E. cuniculi* growing on a pacemaker [8]. This case highlights the potential for antibiotic-resistant endocarditis to be caused by microsporidia. *E. cuniculi* was originally recognized for its ability to disseminate to the nervous system in rabbits, causing seizures. In humans, this species has also been the cause of severe cerebral infections and seizures in immunocompromised patients [9, 10]. In addition to *E. cuniculi*, *Encephalitozoon intestinalis* and *Encephalitozoon hellem* have also been isolated from immunocompromised patients with diarrhea and have been responsible for infection of several other organ systems.

*Enterocytozoon bieneusi*
*Enterocytozoon bieneusi* is probably the most important intestinal pathogen that afflicts AIDS patients, thought to be responsible for 5 to 30% of patients with chronic diarrhea that have CD4 lymphocyte counts below 0.1 X 10^9/liter [11]. These infections can cause death due to persistent, untreatable diarrhea. *E. bieneusi* infects the epithelial cells of the small intestine and can cause villous atrophy, which reduces the absorptive capacity of the intestine [12]. *E. bieneusi* infections do not appear to respond to albendazole, which has been the drug of choice for other microsporidian infections [13, 14]. While there are indications that *E. bieneusi* infections respond to treatment with fumagillin, there are toxic effects from this drug that limit its use. Thus, there is a critical need for better drug treatment options for *E. bieneusi*.

Several studies in Africa have indicated that *E. bieneusi* infection can be quite common in these areas, both in immunocompromised and immunocompetent patients. Estimates of infection vary widely, but were as high as 76% in one study examining HIV+ patients with watery diarrhea [15]. Reported prevalence rates in the 25 studies conducted on patients with HIV infection before the widespread use of active antiretroviral therapy (HAART) (1989–1998) varied between 2% and 70% depending on the symptoms of the population studied and the diagnostic techniques
employed [16-26]. These studies suggest that asymptomatic carriage can occur in immunocompromised patients. Co-infection with different microsporidia or other enteric pathogens can occur. There was no overall trend in these prevalence studies with regard to country of origin or other demographic characteristics. When combined, these studies identified 375 *E. bieneusi* infections among 2400 patients with chronic diarrhea, for a prevalence of 15% in this population. It is clear that since the institution of ART and its associated immune reconstitution the prevalence of diarrhea among AIDS patients has decreased, as has the incidence of microsporidiosis. At the Tulane National Primate Research Center, there is a prevalence for *E. bieneusi* in non-human primates of approximately 35%, based on spore shedding in feces detected by nested PCR (E.S. Didier, unpublished).

**Other microsporidian species**

At least 10 other microsporidian species can cause infections in humans. For example, seven other species (in addition to two *Encephalitozoon* species mentioned above) have been reported to cause eye infections, including corneal infections in immunocompetent hosts. Three species have been reported to infect muscle tissue. One of these is *Brachiola algerae* (originally named *Nosema algerae* and now reclassified as *Annacallia algerae*), which has a broad host range including mosquitoes. *A. algerae* has caused several serious infections in humans, including a fatal muscle infection in an arthritis patient taking immunomodulating drugs [27]. The mode of infection in this patient was possibly via a mosquito bite. This case highlights the significant zoonotic potential of microsporidia.

**Environmental sources and zoonotic potential of microsporidia**

Microsporidian infections have been reported worldwide. The majority of human infections, such as intestinal infections, are thought to be due to ingesting contaminated food or water. *E. bieneusi*, *E. intestinalis* and *Vittaforma cornea* have been found in a variety of water sources, which led microsporidia to be added to the EPA list of waterborne contaminants of concern. Prior to their identification in humans, microsporidia were found in animals, both vertebrates and invertebrates. *E. bieneusi* and *E. cuniculi* have been found in wild, domestic and farm animals. *E. cuniculi* has been isolated from rabbits and dogs while *E. hellem* has been found in a variety of birds. *E. intestinalis* has been found in domestic animals and in gorillas. Thus, microsporidia infections may be derived from a wide variety of environmental sources and there is substantial zoonotic potential.

**Diagnostic challenges and antimicrobial treatments**

Because microsporidia are small (about the size of bacteria) and unculturable on artificial media, diagnosing these infections can be challenging. While PCR tests exist for certain species, these only detect individual species. It would be ideal to develop pan-microsporidia primers that would allow simultaneous detection of multiple species in the same sample. It would also be helpful to investigate microsporidian strain variability in order to determine the rate of false negatives for existing tests. As mentioned above, albendazole has been successful in treating some microsporidian infections, but does not appear to be effective against *E. bieneusi*. Thus, there is a need for further drug development, which would be facilitated by additional sequence data from microsporidia.
3. Microsporidia background

The phylum Microsporidia contains over 1200 species of obligate intracellular pathogens that infect almost all animal phyla. Microsporidia have a unique lifestyle that is particularly well-suited to surviving in hosts and evading host defenses (see Figure 1). Microsporidia survive outside the host as spores protected from the environment by a thick chitinous wall. Inside the spore is an infection apparatus called a polar tube, which undergoes a dramatic eversion upon encounter with a host and pierces the host cell. The polar tube then acts as a syringe to inject spore contents directly into the host cell. The injected sporoplasm then develops into meront stages, which then differentiate into mature spores. The details of this developmental cycle exhibit phenomenal diversity from species to species. Two examples are illustrated below.

![Microsporidia lifecycle](http://www.dpd.cdc.gov/dpdx/HTML/Microsporidiosis.html)

Microsporidia are obligate intracellular pathogens and therefore must be grown in the context of the host cell. While some microsporidia can only be grown in animals, there are tissue culture models for several species, such as the *Encephalitozoon* species, *Vittaforma corneae* and *A. algerae*. So far, it has not been possible to culture *E. bieneusi* in tissue culture cells, which has been a significant impediment to the field. Spores can only be isolated from human patients, monkeys, or immunocompromised rats (although it is difficult to get large numbers from rodents). The species most closely related phylogenetically to *E. bieneusi* is *Nucleospora salmonis*, which infects salmon. Another common microsporidian species studied in fish is *Pseudoloma neurospora*, which infects zebrafish and is a significant problem for zebrafish researchers.

Probably the majority of microsporidian research has been performed in insects. One of the major goals of insect microsporidia research is to develop biocontrol agents for agricultural
pests, as well as medically relevant insects such as mosquitoes. In addition to \( A. \) algerae, \( Edhazardia \) aedis is another mosquito-infecting species that has been a focus of study [28]. Recently, the nematode \( C. \) elegans has emerged as a model for studies of microsporidia. Wild-caught isolates of \( C. \) elegans are infected with a species called \( Nematocida \) parisii [29]. This system provides a convenient whole animal high-throughput screening system for antimicrosporidia drugs, since screens with obligate pathogens require the host to be present. In addition to \( N. \) parisii, \( C. \) elegans can also be infected by the medically relevant species \( E. \) cuniculi (Troemel E.R., unpublished data).

Little is known about immunity against microsporidia. Because AIDS patients are particularly susceptible to infection, T-cells have been implicated in defense. Supporting this idea, athymic mice have been shown to be hypersensitive to \( E. \) cuniculi infection [30]. Almost nothing is known about innate immune control of these pathogens. One very recent report has identified TLR2 as the first receptor to play a role in mediating inflammatory responses to these pathogens [31].

4. Research questions/goals

Very little sequence is available for microsporidia. The first full genome sequence reported is for one strain of \( E. \) cuniculi, which is 2.9 Mb [4]. A second full genome sequence has just been described for the honeybee infecting species \( Nosema \) ceranae [32]. Genomic surveys have been reported for \( Antonospora \) locustae (used as a pesticide for locusts) and \( E. \) bieneusi [16, 33]. ESTs have also been sequenced from \( A. \) locustae and \( E. \) aedis [16, 34]. While these studies have been very informative, the field would greatly benefit from a comprehensive genomic examination of several species to gain more insight into their unique capabilities, their wide host range, and their persistence in various hosts.

We propose to sequence the genomes of microsporidia species in two general classes: Class I species that are known to infect humans and Class II species that infect organisms that provide important models for research. Some of the Class II species also have zoonotic potential. Obtaining sufficiently quantities of pure DNA from obligate pathogens can be difficult. In addition to their medical and agricultural relevance, these species have been chosen because it is possible to procure enough pure DNA to perform the sequencing studies proposed below. The species we have selected are listed in Table 1, and are further described below.

Class I species

\( Encephalitozoon \) species such as \( E. \) cuniculi, \( E. \) intestinalis, and \( E. \) hellem can be grown in tissue culture cells, which greatly facilitates their study. We propose to sequence 3 separate strains of \( E. \) cuniculi, which have been shown to be molecularly distinct and have been isolated from separate hosts [35]. These will be among the first studies to examine the strain variation within a microsporidian species and will provide genomic tools for further study. In particular, these data will enable researchers to examine whether there is recombination and therefore a sexual cycle in these pathogens. In other fungal pathogens, the sexual cycle has been associated with virulence and the ability to infect immunocompetent hosts. Because of their clinical relevance, tiny genomes, and easy accessibility of pure samples, we also propose to sequence \( E. \) intestinalis and \( E. \) hellem.
We propose to sequence *Vittaforma corneae* because it is another human-infecting species that can also be grown in tissue culture. *Brachiola algerae* (now *Anncallia algerae*) is a species that has significant zoonotic potential, having been responsible for muscle infection and death in an immunocompromised patient. The *A. algerae* strain that caused this death has been isolated and can be propagated in tissue culture cells. Sequencing will be performed on this strain, in addition to a strain isolated from mosquitoes, which are the type host for *A. algerae*.

*Enterocytozoon bieneusi* is one of the most medically relevant species, but it has not been possible to culture this species in tissue culture cells. We will sequence two distinct strains of *E. bieneusi* from immunocompromised patients in Uganda, which appear to represent two distinct genotypes. A genome survey has already been performed for one of these strains.

<table>
<thead>
<tr>
<th>Species</th>
<th>Host</th>
<th>Clade*</th>
<th>Rationale</th>
<th>Strains (#)</th>
<th>Source for sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Encephalitozoon cuniculi</em></td>
<td>Human, dog, rabbit</td>
<td>IV</td>
<td>Human pathogen</td>
<td>3</td>
<td>Tissue culture cells (Keeling)</td>
</tr>
<tr>
<td><em>Encephalitozoon intestinalis</em></td>
<td>Human, domestic animals, gorilla</td>
<td>IV</td>
<td>Human pathogen</td>
<td>1</td>
<td>Tissue culture cells (Keeling)</td>
</tr>
<tr>
<td><em>Encephalitozoon hellem</em></td>
<td>Humans, birds</td>
<td>IV</td>
<td>Human pathogen</td>
<td>1</td>
<td>Tissue culture cells (Keeling)</td>
</tr>
<tr>
<td><em>Enterocytozoon bieneusi</em></td>
<td>Humans, pigs, cattle, cats, dogs, chickens, monkeys</td>
<td>IV</td>
<td>Human pathogen</td>
<td>2</td>
<td>Humans (Tzipori, Akiyoshi)</td>
</tr>
<tr>
<td><em>Vittaforma corneae</em></td>
<td>Humans</td>
<td>IV</td>
<td>Human pathogen</td>
<td>1</td>
<td>Tissue culture cells (Didier)</td>
</tr>
<tr>
<td><em>Anncallia algerae</em></td>
<td>Humans, mosquitoes</td>
<td>V</td>
<td>Human pathogen, zoonotic potential</td>
<td>2</td>
<td>Mosquitoes, human (Becnel, Weiss)</td>
</tr>
<tr>
<td><em>Nucleospora salmonis</em></td>
<td>Salmon</td>
<td>IV</td>
<td>Closely related to <em>E. bieneusi</em></td>
<td>1</td>
<td>Salmon (Kent)</td>
</tr>
<tr>
<td><em>Pseudoloma neurophilia</em></td>
<td>Zebrafish</td>
<td>III</td>
<td>Model system pathogen</td>
<td>1</td>
<td>Zebrafish (Kent)</td>
</tr>
<tr>
<td><em>Edhazardia aedis</em></td>
<td>Mosquito</td>
<td>I</td>
<td>Zoonotic potential</td>
<td>1</td>
<td>Mosquitoes (Fast)</td>
</tr>
<tr>
<td><em>Vavraia culicis</em></td>
<td>Mosquito</td>
<td>III</td>
<td>Zoonotic potential</td>
<td>1</td>
<td>Mosquitoes (Becnel)</td>
</tr>
<tr>
<td><em>Nematocida parisii</em></td>
<td>Nematode</td>
<td>II</td>
<td>Model system pathogen</td>
<td>3</td>
<td>Nematodes (Troemel)</td>
</tr>
</tbody>
</table>

Table 1 Proposed microsporidian species for sequencing.
*Clade assignments from ref. [29] and [36].

**Class II species**
We propose to sequence several microsporidian species that have zoonotic potential and/or infect model organisms that provide important and accessible systems for microsporidia study. Among these are *Vavraia culicis* and *Edhazardia aedis*, which infect mosquitoes. *Pseudoloma neurophilia* infects zebrafish and *N. parisii* infects *C. elegans*, which are both genetically tractable hosts that are transparent, allowing for easy visualization of infection. There are three isolates of *N. parisii* that have been isolated from environmentally distinct hosts. We propose to also sequence the salmon-infecting *Nucleospora salmonis*, because it is the closest-described
relative of *E. bieneusi*. Both fish microsporidia are common in their respective hosts, and Dr. Kent (who works actively with both host species) can easily provide the needed material.

### Significance of proposed sequencing studies

With acquisition of the genome sequences from the species described above, we will be able to develop a more complete picture of which genes are common to all microsporidian species, and subgroups such as the human pathogens. We will use this set of sequences to determine which genes are also specific to microsporidia, which will provide tools for better diagnostics and drug development. Identification of these genes will provide targets for development of pan-microsporidia primer sets to aid microsporidian detection with PCR. Proteins made from these genes could provide drug targets for anti-microsporidia drugs.

In addition, these data will provide information that may provide insight into the determinants of virulence and host range for these different organisms. Because we will be sequencing multiple isolates of *E. cuniculi, N. parisii, E. bieneusi* and *A. algerae*, we will also determine the strain variability for these organisms. These comparisons may also address ideas about sexuality in these organisms as it is possible these studies will define mating types. Because genetic manipulation has not been developed for microsporidia, virtually the only method of gene investigation is via misexpression of microsporidian genes in host cells, as has been done for other obligate intracellular pathogens. In order to perform such experiments, genome sequence data are critical. These sequence data will also provide additional insight into the evolutionary origin of these pathogens, which has been controversial. Overall, these data will provide a platform for discovery into an important class of widespread pathogens.

5. Methods/sequencing

#### Sequencing strategy

Our proposed sequencing model for the microsporidian genomes uses 454 single reads to generate contigs, and 454 paired-end reads from a ‘jumping’ library to link these into larger scaffolds. The higher linkage possible with Fosmid (40kb) size clones is not practical for the microsporidia due to the harsh conditions required to isolate DNA from spores. Therefore we will generate 15-fold coverage in 454 fragment reads and 30-fold physical coverage in paired 454 reads from 3kb jumps. If larger jumping libraries are available and the DNA is of sufficient size and quantity, we will consider making >3kb jumps.

We expect that this assembly model will produce a high quality draft assembly of the microsporidia. This model was chosen based on testing different assembly models with bacterial and fungal data. For the fungus *Neurospora crassa*, we generated 454 single ~100 base reads on a GS20 454 instrument and jumping paired-end reads on the FLX 454 instrument. Jumping paired-end reads are ~270 bases in length, containing paired tags of ~110 bases arising from locations separated by ~2.5kb in the genome (paired-end jumping reads). Assemblies with varying levels of coverage of the single and paired-end reads are shown in Figure 2. Assemblies with 15-fold coverage of 100 base fragment reads plus 30-fold physical coverage of paired-end reads yield contigs with an N50 size of ~22kb and a scaffold N50 size of ~530kb. The contigs cover 95% of the reference finished genome sequence, which is comparable to the coverage of the original Sanger assembly. Assemblies with higher read coverage increase the contig size only
slightly (20-fold coverage increase contig size by 15%) so that 15-fold coverage strikes an appropriate balance between cost and contig size. Figure 2 also shows that 30-fold physical coverage of paired-end reads more than doubles N50 scaffold size compared to lower coverage models.

Figure 2. *N. crassa* assembly statistics from 454 fragment and paired-end ‘jumping’ reads showing: A. Contig N50 and B. Scaffold N50, for different levels of standard and paired read assemblies.

Further, assemblies generated with the latest version of Newbler (version 2.0.00) and even longer ~400 base Titanium fragment reads produce even larger contigs with increased N50 size. *E. coli* assemblies (with constant sequence coverage of 15-fold fragment and 30-fold 2.7kb jumping paired-end ‘jumping’ reads) with 100, 270 and 400 base fragment reads yield contig N50 sizes of 50kb, 87kb and 95kb respectively.

Each microsporidian genome will require ½ to one 454 run to sequence. Most have small genomes, including *E. cuniculi* (2.9 Mb), *E. intestinalis* (2.3 Mb), and *E. hellem* (2.5 Mb); *E. bieneusi* is twice the size of these (6 Mb). For the remaining genomes, which do not have an estimated size, we will initially assume a relatively small genome size (~5-6 Mb) for sequencing; if the resulting assembly indicates the genome is substantially larger, we will add additional sequence. The ~6 Mb genomes require 1/4 454 run for both the fragment and jumping libraries. One species on the list, *A. algerae*, is likely to have a substantially larger genome. This genome is estimated to be 15-20 Mb. Assemblies will be assessed by standard quality measures, which include evaluation of sequence depth across the assembly. As recent and large segmental duplications have been observed in some microsporidal genomes, overcollapsing in the assembly can be diagnosed by increases in coverage.

**Quality control of DNA samples to assess host DNA contamination**

Because microsporidia are obligate intracellular pathogens and cannot be grown in the absence of host cells, contamination from host DNA is a significant concern in DNA preparations. In order to gate DNA samples before they are entered into the queue for sequencing at the Broad, we are implementing quality control requirements for all DNA submitted. Several different methods can be used to assess relative amounts of host vs. microsporidian DNA. We are recommending that consortium members use quantitative PCR (qPCR) to assess sample purity prior to submission. Primer sets specific to at least two host genes and two microsporidian genes should be designed and tested to determine primer efficiency. These primers should be at least 90% efficient. These primers can be tested to assess the relative levels of host and pathogen the total DNA sample.
Primers that are 100% efficient will amplify template DNA 1 PCR cycle (1 Ct) earlier in Sample A if it contains twice as much of the target DNA as Sample B. If such primers are used on a sample that contains 90% microsporidian DNA and 10% host DNA, the microsporidian DNA should amplify about 3 Ct’s earlier than host DNA. Comparing template levels with two different primer sets for each template will provide a fairly accurate assessment of the relative contribution of host and microsporidian DNA. If both of these assessments indicate that the DNA prep is at least 50% microsporidia-specific, it can proceed into the sequencing queue at the Broad, with increased sequencing depth as appropriate to provide the target coverage of the microsporidial genome. DNA preps with lower percentages of microsporidian DNA will require evaluation and approval for the amount of sequence required.

6. Community interest
Historically, microsporidian research has focused on species that infect insects and fish hosts. More recently, there has been a surge of interest in human infections with the advent of the AIDS crisis. In addition, there is an increased appreciation of the threat from microsporidia in other immunocompromised patients, as well as immunocompetent individuals. With the technological developments in genomic sequencing and the appreciation of the vector-borne and zoonotic potential of microsporidia, these two groups of researchers are increasingly collaborating to understand these organisms. We have assembled a consortium of researchers from both fields that support the species chosen and will provide samples for their species of interest.

- Donna Akiyoshi Tufts Veterinary School  E. bieneusi
- James Becnel USDA  A. algerae, V. culicis, E. aedis
- Elizabeth Didier Tulane University  Encephalitozoon spp, V. corneae
- Naomi Fast University of British Columbia  E. aedis
- Joe Heitman Duke University  Encephalitozoon spp
- Patrick Keeling University of British Columbia  Encephalitozoons spp
- Michael Kent Oregon State University  P. neurophilia, N. salmonis
- Lee Solter University of Illinois  insect-infecting species
- Emily Troemel University of CA-San Diego  N. parisii
- Saul Tzipori Tufts Veterinary School  E. bieneusi
- Louis M. Weiss Albert Einstein College of Medicine  Encephalitozoon spp, A. algerae
References