

1 **Identification of microsporidia host-exposed proteins reveals a repertoire of large**  
2 **paralogous gene families and rapidly evolving proteins**

3 **Aaron W. Reinke\***, Keir M. Balla, Eric J. Bennett, and Emily R. Troemel

4 Division of Biological Sciences, Section of Cell and Developmental Biology, University of  
5 California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093

6 \*Corresponding author: [awreinke@gmail.com](mailto:awreinke@gmail.com)

7

8 **Abstract**

9 Pathogens use a variety of secreted and surface proteins to interact with and manipulate their  
10 hosts, but a systematic approach for identifying such proteins has been lacking. To identify these  
11 'host-exposed' proteins, we used spatially restricted enzymatic tagging followed by mass  
12 spectrometry analysis of *C. elegans* infected with two species of *Nematocida* microsporidia. We  
13 identified 82 microsporidia proteins inside of intestinal cells, including several pathogen proteins  
14 in the nucleus. These microsporidia proteins are enriched in targeting signals, are rapidly  
15 evolving, and belong to large, *Nematocida*-specific gene families. We also find that large, species-  
16 specific families are common throughout microsporidia species. Our data suggest that the use of  
17 a large number of rapidly evolving species-specific proteins represents a common strategy for  
18 these intracellular pathogens to interact with their hosts. The unbiased method described here for  
19 identifying potential pathogen effectors represents a powerful approach for the study of a broad  
20 range of pathogens.

21

22 **Introduction**

23 Pathogens exploit hosts to promote their own proliferation. Viral, bacterial and eukaryotic  
24 pathogens control their hosts using effector proteins that interact directly with host molecules<sup>1-3</sup>.  
25 These effector proteins can be exported out of the pathogen into host cells or they can remain  
26 attached to the pathogen but with regions of the protein exposed to the host environment. These  
27 host-exposed proteins perform molecular functions that range from manipulation of host defenses

1 to modulation of host pathways that can promote pathogen growth<sup>4,5</sup>. In many cases these  
2 proteins are evolving under diversifying selection, such that variation among these proteins can  
3 influence host survival<sup>6-8</sup>. Examples to date indicate considerable variation in the proteins that  
4 pathogens use to interface with their hosts. The conservation of these host-exposed proteins  
5 varies among different types of pathogens. Whereas most effectors of a strain of *Pseudomonas*  
6 *syringae* are present in other *Pseudomonas* strains and over 35% are conserved in other bacterial  
7 genera<sup>9</sup>, fewer than 15% of predicted host-exposed proteins of *Plasmodium falciparum* are  
8 reported to be conserved among *Plasmodium* species<sup>3</sup>.

9  
10 Comprehensive identification of pathogen proteins that are host-exposed is challenging, because  
11 they need to be distinguished from proteins that are localized inside of pathogen cells. Several  
12 studies have addressed this problem by identifying proteins secreted from pathogens into culture  
13 media<sup>10,11</sup>. However, such studies potentially miss proteins that are only present in the native  
14 context. To circumvent this issue, a recent study chemically labeled proteins inside pathogenic  
15 bacteria and then identified those that were delivered inside of host cells<sup>12</sup>. Although powerful,  
16 this approach requires that a pathogen be both culturable and genetically tractable, and thus it is  
17 not generally applicable to many intracellular pathogens. Additionally, these approaches do not  
18 provide information on the subcellular localization for pathogen proteins within host cells. To  
19 address these limitations, we adapted spatially restricted enzymatic tagging for the study of  
20 pathogen host-exposed proteins. Spatially restricted enzymatic tagging is a recently developed  
21 approach for labeling proteins in specific subcellular locations. This approach uses the enzyme  
22 ascorbate peroxidase (APX) to promote biotin labeling of neighboring proteins, which can be  
23 subsequently purified and identified with mass spectrometry<sup>13</sup>. Here, we take advantage of this  
24 localized proteomics approach to identify host-exposed proteins from microsporidia that are  
25 localized in the intestinal cells of an infected animal.

1 Microsporidia constitute a large phylum of fungal-related obligate intracellular eukaryotic  
2 pathogens. The phylum contains over 1400 described species that infect diverse animals  
3 including nematodes, arthropods, and vertebrates, although individual species often have a  
4 narrow host range<sup>14,15</sup>. Dependent on their hosts for survival and reproduction, they have reduced  
5 genomes that lack several key regulatory and metabolic pathways<sup>16,17</sup>. Together these properties  
6 make microsporidia an excellent model of pathogen evolution. Despite the fact that microsporidia  
7 are of both medical and agricultural importance, tools for genetic modification of microsporidia are  
8 lacking and almost nothing is known about the proteins that enable interactions with their hosts<sup>18</sup>.

9 Two potential targeting signals are known that could expose microsporidia proteins to the host.  
10 These are N-terminal signal-sequences that direct proteins for secretion<sup>19</sup>, and transmembrane  
11 domains that could be used to attach proteins to the pathogen plasma membrane with regions of  
12 the microsporidia protein in direct contact with host molecules<sup>20</sup>. A number of studies have used  
13 these two targeting signals to predict the set of proteins encoded by pathogen genomes that are  
14 likely to be host-exposed<sup>21,22</sup>. However, it is unclear how accurate these approaches are at  
15 identifying such proteins in microsporidia and these prediction methods do not distinguish  
16 between proteins partially or wholly outside the microsporidia cell from those directed to internal  
17 membranes or compartments<sup>13</sup>. Although some host-exposed microsporidia proteins have been  
18 characterized, no comprehensive identification of such proteins has been carried out<sup>23,24</sup>.

19 Several microsporidia of the genus *Nematocida* naturally infect *C. elegans*, a model organism  
20 that offers a number of advantages for the study of host-pathogen interactions<sup>25,26</sup>. Infection of *C.*  
21 *elegans* by *N. parisii* begins with spores being ingested and then invading host intestinal cells. *N.*  
22 *parisii* initially develops in direct contact with the cytoplasm as a meront, eventually differentiating  
23 into a transmissible spore form that exits the cell<sup>27</sup>. Although the infection reduces worm lifespan,  
24 infected animals can generate enormous numbers of spores before death, with a single worm  
25 able to produce over 100,000 spores during the course of the infection<sup>25,28</sup>. Using *C. elegans*, we

1 now report the first unbiased identification of microsporidia host-exposed proteins inside of an  
2 animal. These identified proteins are enriched for rapidly evolving proteins and members of  
3 unique large gene families. We also find that these species-specific large families are common  
4 throughout microsporidia. Using the properties we identified for host-exposed proteins in  
5 *Nematocida*, we analyzed 23 microsporidia genomes to predict potential host-exposed proteins,  
6 almost all were found to have no known molecular function. These results suggest that  
7 microsporidia use a set of lineage specific, rapidly evolving proteins to interact with their hosts.  
8 This study provides a foundation for further functional characterization of host-exposed  
9 microsporidian proteins, and demonstrates the utility of proximity-labeling proteomic methods to  
10 broadly identify pathogen proteins localized within host cells.

## 11 **Results**

### 12 **Experimental Identification of *Nematocida* host-exposed proteins**

13 To identify microsporidia proteins that come into contact with the intracellular host environment  
14 we used the technique of spatially restricted enzymatic tagging<sup>13</sup>. This approach uses the enzyme  
15 ascorbate peroxidase (APX) to label proteins in the compartment where the enzyme is expressed,  
16 with a biotin handle for subsequent purification (Figure 1A). We generated strains of *C. elegans*  
17 expressing GFP-APX, either in the cytoplasm, or in the nucleus of intestinal cells (Figure 1B). We  
18 also generated a negative control strain that expresses GFP in the intestine, but without the APX  
19 protein (Table S1).

20

21 First, we inoculated these transgenic animals with *N. parisii* spores, which led to the majority of  
22 animals being infected (Figure S1). These animals were then incubated for 44 hours at 20°C to  
23 allow for growth of the parasite. Next, we added the biotin-phenol substrate and hydrogen  
24 peroxide to these animals to facilitate APX-mediated biotinylation of host and pathogen proteins  
25 proximal to the GFP-APX protein. Under these conditions we detected biotin-labeled proteins by

1 microscopy in the intestinal cells of infected animals, but no labeling in the microsporidia cells  
2 themselves, demonstrating that the labeling technique is restricted to host-cell regions (Figure  
3 S2). Biotinylated proteins were isolated from total worm extracts using streptavidin-conjugated  
4 resin and these purified proteins were identified using mass spectrometry. Biotinylated proteins  
5 from infected animals were isolated in triplicate and over 4000 proteins from *C. elegans* and *N*  
6 *parisii* were identified (Figure S3).

7 As validation that proteins were labeled in specific compartments in this experiment, we used the  
8 labeled *C. elegans* proteins as an internal control. By comparing spectral counts identified in the  
9 cytoplasmic APX, nuclear APX, and no APX strains, we identified 891 *C. elegans* proteins  
10 specifically labeled in the intestine (Table S2). By comparing *C. elegans* proteins in the  
11 cytoplasmic and nuclear samples we identified 118 proteins specific to the nucleus and 114  
12 proteins specific to the cytoplasm. We then compared these proteins to *C. elegans* proteins with  
13 previously reported localization. The set of proteins we identified as either cytoplasmic or nuclear  
14 specific are enriched for proteins known to be localized in that subcellular compartment (Figure  
15 S4A).

16 Comparing proteins from the cytoplasmic APX and nuclear APX samples to the no APX sample,  
17 we identified 72 *N. parisii* proteins that were enriched above background levels, as defined by the  
18 no APX strain (Table S3). To approximate the total microsporidia proteome detectable in our  
19 experiments, we identified 392 *N. parisii* proteins from the no APX control samples (see methods).  
20 We then compared these protein sets to previously generated RNAseq expression data<sup>22</sup>. The  
21 host-exposed proteins that we identified had moderate mRNA expression levels, with few  
22 detected from either the lowest or highest expressed mRNAs (Figure 2A). In contrast, proteins  
23 identified in the no APX control strain are among the most highly expressed mRNAs in the  
24 genome (Figure S5A). This result suggests that the host-exposed proteins we identified are not  
25 biased towards highly expressed proteins.

1 Compared to all proteins in the genome, the host-exposed proteins we identified were significantly  
2 enriched in both signal peptides and transmembrane domains: over 75% of the proteins identified  
3 (enrichment p-value of 6.6E-13) had at least one predicted targeting signal (Figure 2B). Neither  
4 the proteins identified from the no APX control, nor the identified *C. elegans* intestinal proteins  
5 are enriched for these targeting signals compared to the genome (Figures S4B and S5B).  
6 Altogether, the results indicate that our spatially restricted enzymatic tagging technique identified  
7 a high-quality data set of *N. parisii* host-exposed proteins in *C. elegans*.

8 To investigate the subcellular localization of *N. parisii* host-exposed proteins, we compared  
9 proteins identified from animals expressing APX in the cytoplasm to those identified from animals  
10 expressing APX in the nucleus. From this comparison we found four proteins specific to the  
11 nucleus and eight proteins specific to the cytoplasm. Of the four nuclear specific proteins, three  
12 are predicted to have signal peptides, while all eight cytoplasmic specific proteins are predicted  
13 to have transmembrane domains. These data provide support for a model where proteins with  
14 signal peptides are secreted into the host cell and can localize to different cellular compartments,  
15 including the nucleus. Proteins containing transmembrane domains are likely attached to the  
16 membrane of the pathogen where they come in contact with the host cytoplasm (Figure 2C).

### 17 **Identified *Nematocida* host-exposed proteins are enriched in members of large gene** 18 **families**

19 Large, expanded gene families have been suggested to mediate host-pathogen interactions in a  
20 number of pathogen species and several large gene families have been previously identified in  
21 *Nematocida* species<sup>22,26,29</sup>. We defined large gene families as groups of homologous proteins with  
22 at least ten members in one species that were enriched in signal peptides or transmembrane  
23 domains. We initially identified these families from paralogous orthogroups and then generated  
24 profile hidden Markov models to identify additional members in the genome.

1 There are four large *N. parisii* gene families that contain from 18 to 169 members. Two of these  
2 gene families, NemLGF1 and NemLGF5, encode signal peptides, and the other two gene families,  
3 NemLGF3 and NemLGF4, encode C-terminal transmembrane domains. The host-exposed  
4 proteins we identified are significantly enriched (p-value of 1.3E-16) in these families and contain  
5 35 members of these four genes families, with at least one host-exposed protein in each of the  
6 four families (Figure 2B and C). The four nuclear specific proteins are members of the NemLGF1  
7 or NemLGF5 gene family, whereas four of the cytoplasmic specific proteins with transmembrane  
8 domains belong to the NemLGF3 family (Table S3).

### 9 **Identified *Nematocida* host-exposed proteins are clade specific**

10 To investigate how the repertoire of *N. parisii* host-exposed proteins is evolving, we explored  
11 whether the identified host-exposed proteins are conserved in three other *Nematocida* species.  
12 The earliest known diverging species of the genus is *N. displodere*, which proliferates well in the  
13 epidermis and muscle, but poorly in the intestine<sup>26</sup>. In contrast the other *Nematocida* species are  
14 intestinal-specific<sup>25</sup>. Previously, the species known to be the most closely related to *N. parisii* was  
15 the intestinal-specific *N. sp. 1* (strain ERTm2), which shares 68.3% average amino acid identity  
16 with *N. parisii*. To provide a more closely related species for comparison, we sequenced and  
17 assembled the genome of *Nematocida* strain ERTm5, an intestinal-specific strain that was  
18 isolated from a wild-caught *C. briggsae* in Hawaii<sup>30</sup>. This strain was previously described as a  
19 strain of *N. parisii* based on rRNA sequence, but based on our analysis, it now appears to define  
20 a new species (see methods). This genome is comparable in quality to other sequenced genomes  
21 as judged both by assembly statistics and the presence of proteins conserved throughout  
22 microsporidia (Table S4). This new species, *Nematocida ironsii*, now represents the closest  
23 known sister species to *N. parisii* and has an average amino acid identity of 84.7% compared to  
24 *N. parisii* (Figure S6 and Table S5). To examine conservation, each *N. parisii* protein was placed  
25 into an orthogroup using six eukaryotic and 23 microsporidian genomes. Every *N. parisii* protein

1 was categorized into one of six classes of decreasing conservation: 1) *N. parisii* proteins  
2 conserved with other non-microsporidia eukaryotes, 2) conserved with other microsporidia, 3)  
3 conserved with *N. displodere*, 4) conserved with *N. sp1*, 5) conserved with *N. ironsii*, and 6) those  
4 that are unique to *N. parisii* (Figure 2D).

5 Using this evolutionary approach, we found that the set of host-exposed proteins we identified are  
6 significantly enriched (p-value of 1.9E-20) for less conserved proteins, with only 12% having  
7 orthologs outside of a group of closely related *Nematocida* species (*N. sp. 1*, *N. ironsii* and *N.*  
8 *parisii*, which we refer to as 'clade-specific'). In contrast, 63% of all *N. parisii* proteins in the  
9 genome have orthologs outside of this clade of *Nematocida* species (Figure 2E). Most of these  
10 identified proteins don't have a predicated molecular function, with only five of these 72 proteins  
11 containing a predicted Pfam domain (Figure 2B). To determine the rate of protein evolution, we  
12 calculated the protein sequence divergence between orthologous *N. parisii* and *N. ironsii* proteins.  
13 We found that the host-exposed proteins are rapidly evolving compared to the other proteins in  
14 the genome (Figure 2F).

15 To examine whether the properties of the host-exposed proteins we identified were conserved in  
16 other microsporidia species, we performed spatially restricted enzymatic tagging on *C. elegans*  
17 infected with *N. sp. 1*. Although we identified fewer *C. elegans* and microsporidia proteins from *N.*  
18 *sp. 1* infected animals, we nonetheless found ten proteins enriched over background (Figure S3  
19 and Table S6). These proteins have similar properties to those identified for *N. parisii* as they are  
20 enriched in targeting signals and clade-specific proteins (i.e. proteins not conserved in other  
21 eukaryotes, microsporidia, or *N. displodere*) (Figure S7). They also are enriched for being  
22 members of large gene families, including three members of NemLGF1 and one member of the  
23 *N. sp. 1*-specific family NemLGF6. We also identified two pairs of orthologs from the two species:  
24 hexokinase (NEPG\_02043 and NERG\_02003) and a NemLGF1 family member (NEPG\_02370  
25 and NERG\_01049). To expand this analysis to a different microsporidia genus, we examined data



1 previously generated from germinated *Spraguea lophii* spores. We found that proteins identified  
2 as secreted from these germinated spores were also enriched in the properties of signal peptides  
3 and clade-specific proteins (Figure S8)<sup>24</sup>.

4 Overall, we find that host-exposed proteins are highly enriched in three properties: 1) they have  
5 targeting signals (signal peptides or transmembrane domains), 2) they belong to large gene  
6 families, and 3) they are clade-specific. In fact, 85% of *N. parisii* host-exposed proteins identified  
7 are either members of large gene families, or are clade-specific proteins with a signal peptide or  
8 transmembrane domain (enrichment p-value of 1.7E-25) (Figure 2E). Although the number of  
9 proteins we identified with these properties is 61, the total number of proteins with these properties  
10 encoded by the genome is 713.

11  
12 Current limitations of proteomic methods suggest that this approach will not result in the complete  
13 identification of all host-exposed microsporidia proteins. To estimate the sensitivity of this method  
14 we compared the identified *C. elegans* intestinal proteins to the total number of mRNAs expressed  
15 in the intestine<sup>31</sup>. We also compared the total number of detected *N. parisii* proteins to the number  
16 encoded by the proteome. From these comparisons we estimate that we identified between ~8-  
17 24% of potential host-exposed proteins. This would mean that the total host-exposed proteome  
18 encoded by *N. parisii* is on the order of 300 - 900 proteins, a range that encompasses the number  
19 of proteins in the genome that have the properties enriched in the experimentally identified host-  
20 exposed proteins.

21  
22 **Large families display lineage specific expansions and are common in microsporidia**

23 If most members of *N. parisii* large gene families are involved in host-pathogen interactions, we  
24 would predict that they would also be rapidly evolving with species-specific radiations. The four  
25 large gene families of *N. parisii* contain a total of 295 members. Members of these four families

1 are also present in the other species in this clade, *N. sp. 1* and *N. ironsii*, but not any other  
2 microsporidia species (Figure 3A and Figure 4). Phylogenetic trees of these families show  
3 expansion of family members specific to each species (Figure 3B and C). Members from these  
4 families are often not conserved between species, with only 5-39% of *N. parisii* members in each  
5 gene family that have orthologs in *N. sp. 1* and 56-95% that have orthologs in *N. ironsii* (Figure  
6 3C). The largest families that have signal peptides, NemLGF1 and NemLGF5, are enriched for  
7 genes on the ends of chromosomes, a chromosomal localization that is not enriched in the  
8 transmembrane-containing families (Figure S9A). The four families are often adjacent to each  
9 other, suggesting they are being generated through local duplication events (Figure S9B).

10 To examine whether large gene families are common in other microsporidia species, we  
11 examined 23 microsporidia genomes (17 other microsporidia species and six from *Nematocida*)  
12 (Figure S6). From these 21 species, 68 families were identified with at least ten members in one  
13 species and enriched in either predicted signal peptides or transmembrane domains. In addition,  
14 we found that most (59 of 68) of these families do not have any members present outside of the  
15 genus or species. For example, there are three families with members present in all four  
16 *Encephalitozoon* species but no other species examined. Additionally, we identified four large  
17 gene families that were conserved throughout most microsporidia including two ricinB domain  
18 containing families<sup>24</sup>. All but one species examined has a large genus-specific family,  
19 demonstrating that large gene families are widespread throughout microsporidia.

## 20 **Prediction of putative host-exposed proteins from other microsporidia genomes**

21 We next investigated whether proteins that are not widely conserved in microsporidia share  
22 properties with the identified host-exposed proteins. We examined 23 microsporidian genomes to  
23 identify proteins that are not conserved with other eukaryotes, or conserved with distantly related  
24 microsporidia species. These clade-specific proteins are all significantly enriched in targeting  
25 signals compared to proteins conserved with more distally related microsporidia or other

1 eukaryotes (Figure 5A). This result is similar to what we found in our analysis of experimentally  
2 identified host-exposed proteins in *Nematocida*, and similar to a previous study of several  
3 microsporidian species<sup>32</sup>.

4  
5 Our analyses above indicated that the genomes of microsporidia contain two classes of proteins  
6 enriched in targeting signals, clade-specific proteins and large gene families. Most of the proteins  
7 (85%) we identified experimentally in *N. parisii* also display these characteristics. Based on these  
8 genomic signatures and our experimental results, putative host-exposed proteins for each species  
9 were predicted. These predictions of 11,675 proteins for 23 genomes are provided as a resource  
10 in Table S7. Although these characteristics alone may not be sufficient to direct proteins to  
11 become host exposed, these proteins likely represent a substantial portion of the host-exposed  
12 proteins that each species uses and provide an unprecedented set of candidates for future  
13 studies.

14 The potential host-exposed proteins account for 6-32% of the genome of each species.  
15 Interestingly, the number of predicted host-exposed proteins can vary even within closely related  
16 species, with *E. cuniculi* having almost twice as many predicted proteins as the other members  
17 of the genus (Figure 5B). The majority of these putative host-exposed proteins do not have a  
18 predicted molecular function, with only 7.4% having a predicted Pfam domain that occurs in  
19 proteins outside of microsporidia (Table S7). Although most of these proteins do not have known  
20 domains, several species have expanded families of leucine rich repeat (LRR) domains and two  
21 species have expanded families of protein kinases (Figure 5). The most frequently observed  
22 domains in putative host-exposed proteins that are not members of the large gene families are  
23 transporters, kinases, LRR domains, ubiquitin carboxyl-terminal hydrolases, and the bacterial  
24 specific DUF1510 (Figure S10)<sup>33</sup>. Interestingly, a number of domains that are present in the large  
25 gene families are also observed in the non-paralogous proteins, suggesting that there are several

1 common domains that have been utilized in multiple microsporidia species to interact with hosts.  
2 These predictions of host-exposed proteins suggest that microsporidia employ a large number of  
3 proteins with novel domains to interact with hosts.

#### 4 **Discussion**

5 To understand how microsporidia interact with their hosts, we experimentally identified 82 host-  
6 exposed proteins from two *Nematocida* species. To identify these proteins, we employed an  
7 unbiased approach that labeled the host-exposed pathogen proteins inside of an intact animal.  
8 Attempts to validate these host-exposed proteins using orthogonal experimental approaches  
9 have not been possible due to our inability to raise specific antibodies against *Nematocida*  
10 proteins and the lack of genomic modification techniques for microsporidia<sup>34</sup>. Nonetheless, this  
11 approach was able to identify *C. elegans* proteins previously shown to be localized to the nucleus  
12 and cytoplasm, validating the specificity of the technique. This approach of tagging pathogen  
13 proteins based on their localization is likely to be useful in the study other *C. elegans* pathogens  
14 as well as a general tool to examine putative pathogen effector proteins in a range of hosts.

15  
16 A key feature of the identified host-exposed proteins is their enrichment in signal peptides and  
17 transmembrane domains. This enrichment suggests that these are the two major targeting signals  
18 that are used in *Nematocida* for proteins to become exposed to the host, as they are present in  
19 76% of identified proteins. Such signals might be missed in the remaining proteins due to the lack  
20 of sensitivity of these prediction methods and the misannotation of the true N- and C- termini of  
21 *Nematocida* proteins<sup>19,20</sup>. The identified proteins could also be useful to discover potential  
22 secondary signals in the proteins that direct transmembrane and signal peptide containing  
23 proteins to become host exposed, rather than to other membranes inside microsporidia<sup>35</sup>.

24 We found that large gene families are common within microsporidia, with 68 gene families from  
25 23 microsporidia genomes being identified. Although several of these families had been

1 previously reported, here we provide a comprehensive identification of these gene families  
2 throughout microsporidia<sup>24,26,36,37</sup>. The majority of these large gene families have no known  
3 molecular function based on sequence similarity. One enticing possibility is that the expansion of  
4 these families is due to interactions with host proteins. In support of this possibility, a number of  
5 the gene families with predicted domains are known to mediate protein-protein interactions  
6 including LRR and RING domains.

7 One intriguing characteristic of these large gene families is that they are either genus- or species-  
8 specific, with large lineage specific expansions of these gene families across microsporidia. The  
9 differences in the total number of gene families can be quite large in the same genus. For  
10 example, in the family NemLGF3, *N. sp. 1* only has three members compared to 53 members in  
11 *N. parisii*. Both strains of *N. sp. 1* also have a gene family (NemLGF6) that is absent in other  
12 microsporidia, but the ERTm2 strain of *N. sp. 1* has 23 members of a multitransmembrane gene  
13 family (NemLGF7) that is absent from the ERTm6 strain (Figure 4). These differences suggest  
14 that the composition and emergence of these gene families can change rapidly.

15 Several of the large gene families in *Nematocida* contain over 100 members and constitute a  
16 sizable portion of the entire genome. For example, members of NemLGF1 account for 6.4% of  
17 the genome of *N. parisii* and members of NemLGF2 account for 10.8% of the *N. displodere*  
18 genome. The exact forces that are providing pressure for gene family expansion in microsporidia  
19 are unclear, though one likely possibility is that variation in the host environment shapes the  
20 expansion of pathogen protein families. In the case of *N. displodere*, the pathogen has been  
21 observed to replicate in multiple tissues, and this variation in cellular environment could drive  
22 family diversification. Another possibility is that the genetic diversity of the hosts being  
23 encountered could drive the expansion. The complete native ecology of hosts that *Nematocida*  
24 interact with is unknown, though both *C. elegans* and *C. briggsae* have been found infected with  
25 *Nematocida* microsporidia<sup>25</sup>. For other microsporidia species there is both ecological evidence

1 and laboratory studies demonstrating that the same strain of microsporidia can infect closely  
2 related host species<sup>15,38,39</sup>. We speculate this host diversity could drive the expansion of large  
3 gene families in microsporidia and that these large gene families may in turn influence the host  
4 range.

5 The majority of the host-exposed proteins we identified in *N. parisii* and *N. sp. 1* were proteins not  
6 conserved with *N. displodere* or other microsporidia species. Although lack of conservation  
7 accounts for most of the proteins identified, several conserved proteins were identified, including  
8 hexokinase, which we identified in both *Nematocida* species. Hexokinase was previously found  
9 to have predicted signal peptides in several microsporidia species and to be secreted from the  
10 microsporidia *Antonospora locustae*, providing experimental evidence that secreted hexokinase  
11 is a conserved feature of microsporidia<sup>22,23,32</sup>. There are also several large gene families that have  
12 members present in multiple microsporidia species. This observation suggests that although  
13 selective forces result in a host-exposed protein repertoire with many unique proteins for each  
14 microsporidia clade, there are some proteins conserved throughout microsporidia involved in host  
15 interactions.

16 A number of forces are likely to shape the repertoire of host-exposed proteins, including the  
17 selective pressure of the host and interactions with other pathogens. Many of the features of the  
18 host-exposed protein repertoire in microsporidia are similar to characteristics reported in the  
19 apicomplexan phylum of protozoan obligate intracellular pathogens. Large gene families with  
20 either signal peptides or transmembrane domains are common. These families often have  
21 subtelomeric genomic locations and are species specific<sup>29</sup>. Over 200 secreted proteins have been  
22 predicted in *P. falciparum* and few are conserved with other *Plasmodium* species<sup>3</sup>. Most of these  
23 proteins also have no predicted molecular function<sup>35</sup>. These similarities among species suggest  
24 that similar selective pressures can sculpt a host-exposed protein repertoire with related  
25 properties. In contrast, strains of the bacteria *P. syringae* are predicted to have less than 40 type

1 III effectors and contain many effectors shared with other bacteria, many of these which display  
2 evidence of horizontal gene transfer<sup>9,40</sup>.

3 A striking result of our analysis is that a large number of experimentally identified and predicted  
4 host-exposed proteins do not have domains found outside of microsporidia. These host-exposed  
5 proteins are a potential source of novel biochemical activity as the extreme selective pressures  
6 inflicted on pathogens by the host has been shown to result in unique molecular functions<sup>41,42</sup>.  
7 Interestingly, we also predict a large percent of the microsporidia genome to be responsible for  
8 mediating host-pathogen interactions. This suggests that although microsporidia have the  
9 smallest known genomes of any eukaryotes they somewhat paradoxically encode a substantial  
10 cadre of proteins for interacting with their hosts. Understanding how microsporidia use these  
11 proteins to mediate host-interactions will provide insight into their impact on hosts and the  
12 constraints on evolution of a minimalistic eukaryotic genome.

13

#### 14 **Acknowledgements**

15 We thank Steven Wasserman, Matthew Daugherty, and members of the Troemel lab for providing  
16 helpful comments on the manuscript. AWR is a Monsanto Fellow of the Life Sciences Research  
17 Foundation.

18

#### 19 **Author Contributions**

20 AWR designed, conducted, and analyzed experiments and co-wrote the paper. KMB provided  
21 the *N. ironsii* genome sequence. EJB performed the mass spectrometry analysis and co-wrote  
22 paper. ERT provided mentorship and co-wrote the paper.

#### 23 **Competing financial interests**

24 The authors declare no competing financial interests.

25

## 1 **Materials and Methods**

2

### 3 **Cloning and generation of *C. elegans* expressing APX**

4 Soybean APX (W41F) was optimized for *C. elegans* expression using DNAsworks to design  
5 primers<sup>43</sup>. These primers were annealed using a two-step PCR method and cloned into Gateway  
6 plasmid pDONR 221. Gibson cloning was then used to introduce GFP as an N-terminal fusion,  
7 and NES (LQLPPLERLTLD) and NLS (PKKKRKVDPKKRKVDPKKRKV) tags to the C-  
8 terminus of APX<sup>44</sup>. 1 kilobase (kb) of sequence upstream of the intestinal-specific gene *spp-5* was  
9 used as a promoter and *unc-54* as a 3 prime sequence. Multisite Gateway was used to combine  
10 these fragments into the plasmid pCFJ150 to generate targeting constructs. The MosSCI  
11 approach was used to generate single copy insertions by injecting *unc-119* mutants from the  
12 EG6699 strain with these targeting constructs<sup>45</sup>. Each transgenic strain was backcrossed to the  
13 wild-type N2 strain 3 times and the homozygote was used in subsequent experiments.

14

### 15 **Spatially restricted enzymatic tagging of microsporidia infected *C. elegans***

16 *C. elegans* strains that express GFP-APX either localized to the cytoplasm or nucleus, as well as  
17 a control GFP only strain were used (Table S1). Mixed-stage populations of each strain were  
18 grown at 20°C on nematode growth media (NGM) plates seeded with OP50-1 bacteria. Animals  
19 were washed off of plates with M9 and treated with sodium hypochlorite solution/1M NaOH for 2-  
20 3 minutes. Eggs were washed 3 times with M9 and resuspended in 5ml of M9 in a 15 ml tube.  
21 These eggs were incubated 18-24 hours at 20°C on a rotator to hatch L1 animals. Animals were  
22 infected with microsporidia, *N. parisii* (strain ERTm1) and *N. sp. 1* (strain ERTm2), using spores  
23 that were purified as previously described<sup>25</sup>. Infections were performed in 15 ml tubes containing  
24 ~150,000 L1s in 500 µl M9 and 10 µl of 10X concentrated OP50 bacteria, to which 405 µl of *N.*  
25 *parisii* (44.45X10<sup>6</sup> spores) or *N. sp. 1* (14.85 X10<sup>6</sup> spores) were added. These animals were  
26 incubated with spores for 4 hours at 20°C. Animals were then washed 3 times with M9. Animals



1 were resuspended in 12.5 ml M9 and 2.5 ml was added to each 15 cm RNAi plates seeded with  
2 HT115 bacteria expressing *bus-8* RNAi feeding clone<sup>46</sup>. This RNAi clone increases permeability  
3 of the cuticle and allows for efficient biotin labeling<sup>47</sup>. Infected animals were grown for 44 hours at  
4 20°C. Animals were recovered off of each plate with M9T (M9/0.1% Tween-20) and animals  
5 washed once with M9T. To worms in a total of 100 µl M9T in 1.5 ml tubes, 900 µl of labeling  
6 solution (0.1% Tween-20, M9, 3.3 mM biotin-phenol, synthesized as previously described<sup>13</sup>) was  
7 added. Worms were incubated for 1 hour at 22-24°C on a rotator. Then 10 µl of 100 mM H<sub>2</sub>O<sub>2</sub> was  
8 added for 2 minutes. The reaction was quenched with 500 µl quench buffer (M9/0.1% TWEEN-  
9 20/ 10 mM sodium azide/ 10 mM sodium ascorbate/ 5mM Trolox). Samples were washed 4 times  
10 with 1 ml with quench buffer. To each worm pellet 800 µl lysis buffer (150 mM NaCl/ 50 mM TRIS  
11 pH 8.0/1% TritonX-100/0.5% Sodium deoxycholate/0.1%SDS/10 mM sodium azide/ protease  
12 complete tablet (Roche)/ 10 mM sodium ascorbate/ 5mM Trolox/1 mM PMSF) was added and  
13 worms were then immediately frozen dropwise in liquid N<sub>2</sub>.

14  
15 Frozen worm pellets were ground to a fine powder in liquid N<sub>2</sub> to generate protein extracts. These  
16 protein extracts were then centrifuged for 10 min 21,000 g at 4°C. The supernatant was then  
17 filtered over a desalting column (Pierce). The protein concentrations of the extracts were  
18 normalized using a Pierce 660nm Protein Assay. To 340 µg of each sample was added 25 µl of  
19 high capacity streptavidin agarose resin (Pierce) in a total of 700 µl lysis buffer. Extracts were  
20 incubated with beads for 1 hour on rotator. Beads were then washed 5 times with 1 ml lysis buffer,  
21 3 times with 1 ml 8M urea/10 mM TRIS pH 8.0, and 3 times with 1 ml PBS. The liquid was removed  
22 from the beads and 100 µl of 0.1 µg/µl trypsin (Promega)/50 mM NaHCO<sub>3</sub> was added to each  
23 sample and incubated at 37°C for 24 hours.

24

25 **LC-MS-MS parameters**

1 Samples were analyzed in triplicate by LC-MS/MS using a Q-Exactive mass spectrometer  
2 (Thermo Scientific, San Jose, CA) with the following conditions. The following is a generalized  
3 nHPLC and instrument method that is representative of individual analyses. Peptides were first  
4 separated by reverse-phase chromatography using a fused silica microcapillary column (100  $\mu$ m  
5 ID, 18 cm) packed with C18 reverse-phase resin using an in-line nano-flow EASY-nLC 1000  
6 UHPLC (Thermo Scientific). Peptides were eluted over a 100 minute 2-30% ACN gradient,  
7 followed by a 5 minute 30-60% ACN gradient, a 5 minute 60%- 95% gradient, with a final 10  
8 minute isocratic step at 0% ACN for a total run time of 120 minutes at a flow rate of 250 nl/ min.  
9 All gradient mobile phases contained 0.1% formic acid. MS/MS data were collected in a data-  
10 dependent fashion using a top 10 method with a full MS mass range from 400-1800 m/z, 70,000  
11 resolution, and an AGC target of 3e6. MS2 scans were triggered when an ion intensity threshold  
12 of 1e5 was reached with a maximum injection time of 60ms. Peptides were fragmented using a  
13 normalized collision energy setting of 25. A dynamic exclusion time of 40 seconds was used and  
14 the peptide match setting was disabled. Singly charged ions, charge states above 8 and  
15 unassigned charge states were excluded.

16

### 17 **Peptide and protein identification and quantification**

18 The resultant RAW files were converted into mzXML format using the ReadW.exe program. The  
19 SEQUEST search algorithm (version 28) was used to search MS/MS spectra against a  
20 concatenated target-decoy database comprised of forward and reversed sequences from the  
21 reviewed UniprotKB/Swiss-Prot FASTA *C. elegans* database combined with the UniprotKB *E. coli*  
22 (K12 strain) database, and the *N. parisii* and *N. sp. 1* predicted proteomes with common  
23 contaminants appended. The search parameters used are as follows: 20 parts per million (ppm)  
24 precursor ion tolerance and 0.01 Da fragment ion tolerance; up to three missed cleavages were  
25 allowed; dynamic modification of 15.99491 Da on methionine (oxidation). Peptide matches were  
26 filtered to a peptide false discovery rate of 2% using the linear discriminant analysis (Huttlin et al.,

1 2010). Proteins were then filtered to a 2% false discovery rate (FDR), which resulted in a peptide  
2 FDR below 1%. Peptides were assembled into proteins using maximum parsimony and only  
3 unique and razor peptides were retained for subsequent analysis. Peptide spectral count data  
4 was mapped onto the assembled proteins and used for subsequent analysis.

5

### 6 **Analysis of mass spectrometry data**

7 The peptide spectral counts of proteins were used to calculate fold change ratios and FDR p-  
8 values between GFP only, NES, and NLS samples using the qspec-param program of  
9 qprot\_v1.3.3<sup>48</sup>. Several criteria were used to classify proteins as being host-exposed proteins; No  
10 counts in the GFP only samples and an average greater than 2 peptides in the NES samples or  
11 an NES/GFP ratio greater than 2-fold with an FDR p-value of less than 0.005. Additionally proteins  
12 with an NLS/GFP ratio of greater than 3-fold were included. Proteins were classified as being  
13 NLS-enriched if they had a greater than a 2-fold NLS/NES ratio and NLS depleted if they had  
14 greater than a four-fold NES/NLS ratio. All data for *N. parisii* proteins is in Table S3 and for *N.*  
15 *sp. 1* proteins in Table S6. *C. elegans* intestinal proteins were detected in the same way as  
16 described above and data are in Table S2. *N. parisii* proteins in the no APX sample were required  
17 to have an average of greater than 2 peptides in the GFP only sample.

18

### 19 **Microscopy of infected *C. elegans***

20 To detect biotin labeling in infected worms, intestines were dissected and stained with anti-GFP  
21 (Roche) and Streptavidin Alexafluor 568 (Thermo Fisher). Images were taken using a Zeiss  
22 LSM700 confocal microscope with a 40x objective. To detect microsporidia in infected worms,  
23 fluorescence in situ hybridization with probes specific for microsporidia was performed as  
24 previously described and imaged with a Zeiss AxioImager M1 microscope<sup>49</sup>.

25

### 26 **Genome sequencing and analysis**

1 Genomic DNA was obtained from *Nematocida* strain ERTm5 infected animals by phenol-  
2 chloroform extraction, treated with RNase for one hour, and then precipitated with ethanol and  
3 resuspended in TE buffer. One lane of 100 bp paired-end sequencing on an Illumina HiSeq 2000  
4 (Cofactor Genomics) was used to generate reads which were filtered to remove *C. elegans* and  
5 *E. coli* genome reads.

6  
7 The genome was assembled and annotated as done previously<sup>26</sup>. Although ERTm5 was  
8 previously considered to be a strain of *N. parisii* based on 100% nucleotide identity of 18S  
9 ribosomal RNA sequences<sup>30</sup>, the average nucleotide identity across the genome between *N.*  
10 *parisii* strain ERTm1 and ERTm5 is 92.3%, which was calculated using the nucmer program in  
11 mummer 3.23<sup>50</sup>. The two strains are more dissimilar than the generally accepted definition of  
12 different microbial species having less than 95% average nucleotide identity<sup>51</sup>. Because of this  
13 we consider strain ERTm5 to be a separate *Nematocida* species. Because the strain ERTm5 was  
14 isolated from Kauai, Hawaii we named this new species *Nematocida ironsii* in dedication to the  
15 Hawaiian surfer Andy Irons. This Whole Genome Shotgun project has been deposited at  
16 DDBJ/ENA/GenBank under the accession LTKD00000000. The version described in this paper  
17 is version LTKD01000000. Assembly statistics for all microsporidia species used in this study are  
18 in Table S4. Annotation of *N. ironsii* proteins are in Table S5. Conservation of proteins for each  
19 microsporidia species was determined by counting the number of orthogroups conserved  
20 between all 23 genomes divided by the number of orthogroups present in the other species. A  
21 phylogenetic tree of the microsporidia species was generated as described previously (Figure  
22 S6)<sup>26</sup>.

23

#### 24 **Functional annotation of microsporidia proteins**

25 Domains were predicted with the Pfam-A 28.0 library using the HMMscan function in HMMER 3.1  
26 with an E-value cutoff of less than  $10^{-5}$ . Prediction of signal peptides was done using SignalP 4.1,

1 using the best model with a cutoff of 0.34 for both the noTM model and for the TM model<sup>19</sup>.

2 Prediction of transmembrane domains was done using TMHMM 2.0<sup>20</sup>.

3

#### 4 **Determination of microsporidia orthogroups**

5 Conservation of proteins was determined using OrthoMCL 2.0.9<sup>52</sup>. This analysis was performed

6 using six eukaryotic genomes (*Saccharomyces cerevisiae*, *Monosiga brevicollis*, *Rozella*

7 *allomycis*, *Neurospora crassa*, *Ustilago maydis*, and *Allomyces macrogynus*) and 23

8 microsporidia genomes (Table S4) using an inflation index of 1.5 and a BLAST e-value cutoff of

9  $10^{-5}$ .

10

#### 11 **Identification of large gene families in microsporidia genomes**

12 Families were initially identified from microsporidia orthogroups. Proteins in each initial group

13 were aligned using MUSCLE 3.8.31<sup>53</sup> and profile HMM models were built using HMMbuild. The

14 microsporidia genomes were then searched using HMMscan with an E-value cutoff of  $10^{-5}$ . This

15 process was performed iteratively until no more additional proteins met the cutoff. The following

16 domains are widely present in eukaryotic species so family membership was determined using

17 orthogroups: LRR, kinase, ABC transporter, peptidase, and chitin synthase. To be considered a

18 large gene family at least 10 unique proteins had to belong to the family in a single microsporidia

19 genome. Additionally, each family was required to have at least 2-fold enrichment over the

20 genome in either predicted signal peptides or transmembrane domains. Families were named by

21 first three letters of genus and numbered based on size. Those that were present in multiple

22 genera were named with the prefix "Mic".

23

#### 24 **Determination of *N. parisii* large gene family orthologs**

25 For each of four large gene families (NemLGF1 and NemLGF2-4) members from *N. parisii*

26 (ERTm1 and ERTm3), *N. ironsii*, and *N. sp. 1* (ERTm2 and ERTm6) were aligned using MUSCLE.

1 Phylogenetic trees were inferred for each family using RAxML 8.2.4<sup>54</sup> using the PROTGAMMALG  
2 model and 1000 bootstrap replicates. For NemLGF1, an initial tree was generated using 10  
3 bootstrap replicates and then divided into 7 sub trees. Orthologs of *N. parisii* proteins in each  
4 family were manually assigned using these maximum likelihood trees. To determine the genomic  
5 location of these families the 5 largest scaffolds of *N. parisii* (ERTm1) were used. Chromosomal  
6 ends were defined as the first and last 30 kb of each scaffold. Adjacent proteins were calculated  
7 as where the next protein was next to it.

8

### 9 **Determination of conservation**

10 For *N. parisii* proteins, conservation was determined based on orthogroups, except for the large  
11 gene families NemLGF1 and NemLGF2-4 for which orthology was determined as described  
12 above. The following procedure was used to place the *N. parisii* proteins into 6 categories. If a *N.*  
13 *parisii* protein was in any group with a protein from the 6 non-microsporidian eukaryotic species,  
14 the protein was placed in the category “Eukaryotes”. If any remaining unassigned proteins were  
15 in a group with a protein from the microsporidia species not in the genus *Nematocida*, then it was  
16 placed in the category “microsporidia”. If any remaining unassigned proteins were in a group with  
17 an *N. displodere* protein, then it was placed in the category “*N. displodere*”. If any remaining  
18 unassigned proteins were in a group with an *N. sp. 1* protein, then it was placed in the category  
19 “*N. sp. 1*”. If any remaining unassigned proteins were in a group with a *N. ironsii* protein, it was  
20 placed in the category “*N. ironsii*”. The remaining proteins were placed in the category “*N. parisii*”.

21

22 To predict host-exposed proteins the conservation of microsporidia proteins was determined.  
23 Proteins of each species were placed into two classes, “Conserved” or “clade-specific”. If a protein  
24 was in the same group as a protein from any of the eukaryotic or microsporidia species then it  
25 was classified as “conserved”. Otherwise it was classified as “clade-specific”. This was done  
26 except for the closer related species where proteins in the same clade were not considered. For

1 this purpose the following clade definitions were used: *Nematocida* species are *N. parisii*, *N. sp.*  
2 1, and *N. ironsii*; *Encephalitozoon* species are *E. romaleae*, *E. hellem*, *E. intestinalis*, *E. cuniculi*,  
3 and *O. colligate*; and the species *V. culicis* and *T. hominis*.

4

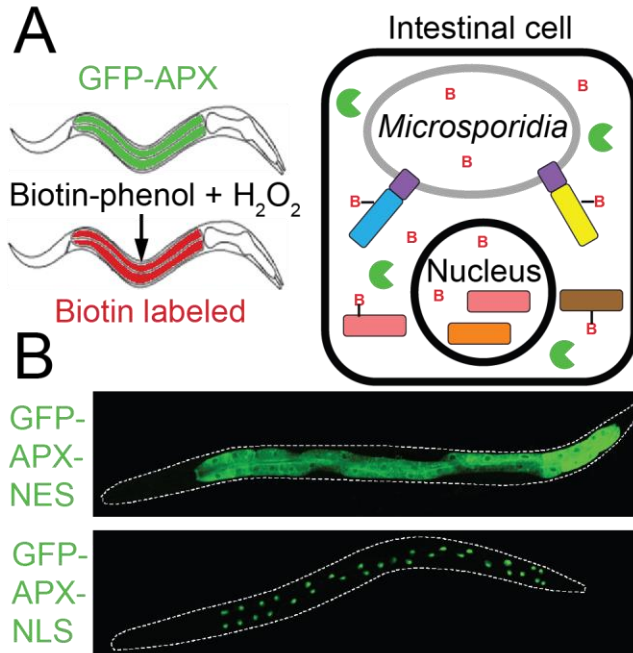
#### 5 **Calculation of protein sequence divergence.**

6 Proteins for microsporidia genomes were placed into orthogroups as described above. Proteins  
7 from one-to-one orthologs of the two *N. parisii* strains (ERTm1 and ERTm3) and *N. ironsii* were  
8 aligned using MUSCLE 3.8.31<sup>53</sup>. For large gene families orthologs were determined as described  
9 above. For proteins conserved with *N. sp. 1*, the evolution rate was only calculated for one to one  
10 orthologs between the 5 genomes. For proteins conserved with *N. displodere*, the evolution rate  
11 was only calculated for one-to one orthologs between all 6 *Nematocida* genomes. Maximum  
12 likelihood trees were built using ortholog sets (three sequences per set) of aligned protein  
13 sequences using PHYLIP (<http://evolution.genetics.washington.edu/phylip.html>). The sum of the  
14 sequence tree length divided by the number of sequences, in PAM units, was calculated for each  
15 ortholog set.

16

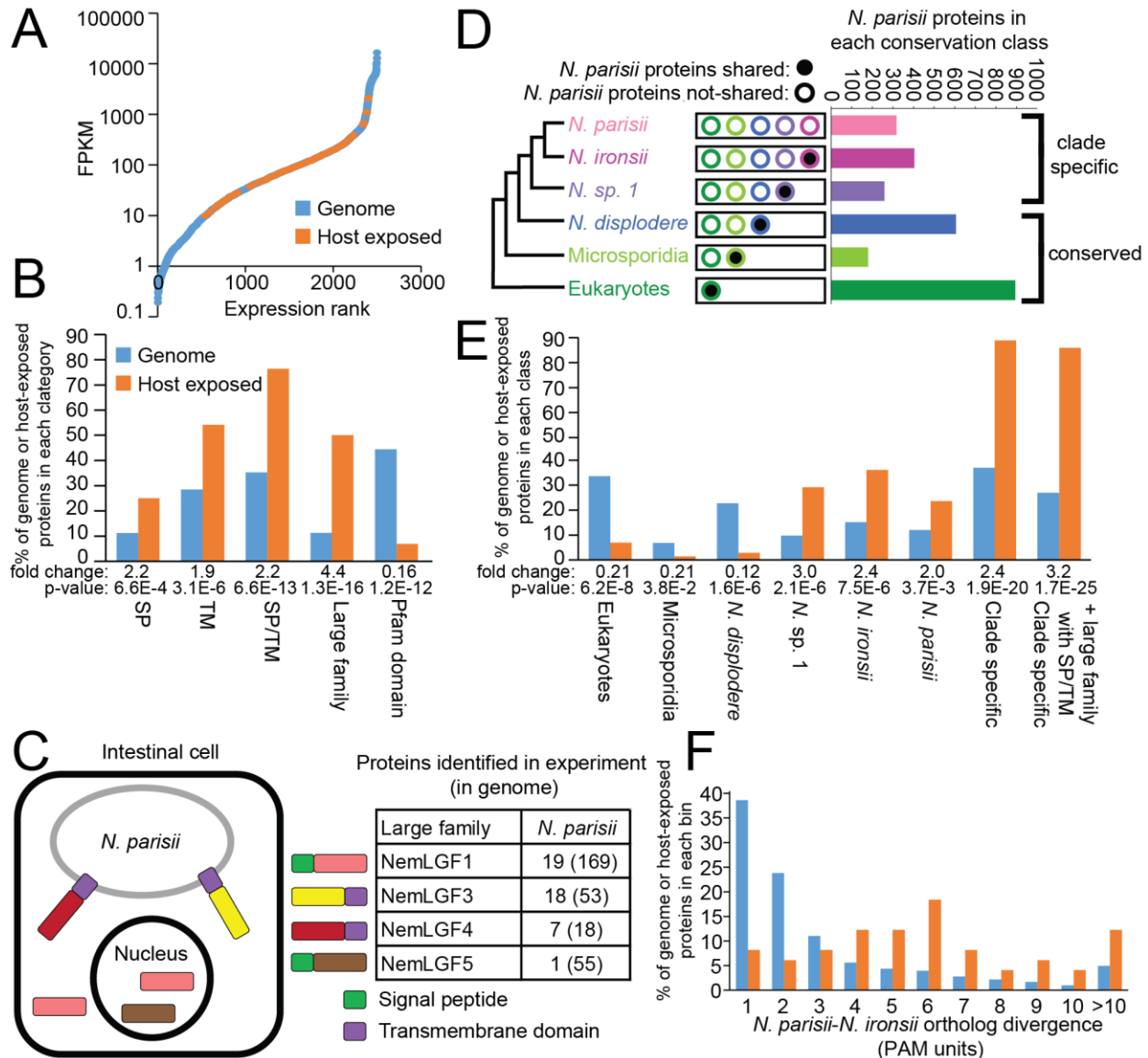
#### 17 **Figures**

18



1  
2 **Figure 1. Overview of approach to detect and analyze host-exposed microsporidia**  
3 **proteins. A.** Schematic of spatially restricted enzymatic tagging in *C. elegans*. Left, worms  
4 expressing GFP-APX in the cytoplasm of the intestine and infected with microsporidia are treated  
5 with biotin-phenol and H<sub>2</sub>O<sub>2</sub>. This treatment results in proteins within the intestinal cytoplasm  
6 being labeled with biotin. Right, an intestinal cell infected with microsporidia expressing  
7 cytoplasmic APX (green circular sectors) labeling microsporidia host-exposed proteins with biotin  
8 (red B). **B.** Animals expressing GFP-APX in the intestine localized to either the cytoplasm (top)  
9 or the nucleus (bottom).





1  
2 **Figure 2. Properties of experimentally identified *N. parisii* host-exposed proteins. A.**  
3 Comparison of mRNA expression levels of identified host-exposed proteins (orange dots) to the  
4 rest of the expressed *N. parisii* proteins (blue dots). Expression data are from a previous RNA-  
5 seq study on animals infected for 30 hours at 25°C<sup>49</sup>. **B, E.** Comparison of identified host-exposed  
6 proteins (orange) to the genome (blue). Enrichment fold change and p-values (one-side Fisher's  
7 exact test) of the host-exposed proteins compared to the genome are listed below each category.  
8 **B.** Properties of 72 *N. parisii* host-exposed proteins. The percentage of the *N. parisii* genome and  
9 the percentage of the host-exposed proteins in each category are shown. TM, transmembrane.  
10 SP, signal peptide. **C.** Left, model of where identified large gene family proteins are localized.  
11 Right, the number of proteins of each gene family identified as host exposed and the total number  
12 of gene family members present in the genome is shown in parentheses. **D.** Schematic of the  
13 categorization of *N. parisii* proteins by conservation class. The 2661 proteins in the genome were  
14 placed into 6 classes of decreasing conservation from proteins conserved with eukaryotes to  
15 being proteins unique to *N. parisii*. **E.** Percentage of the genome and host-exposed proteins in  
16 each conservation class. **F.** Distribution of protein sequence divergence between *N. parisii* and

1 *N. ironsii* one-to-one orthologs. The genome contains 2083 orthologs that met our criteria and the  
2 host-exposed proteins contain 49 orthologs (See methods). The percentage of the identified host-  
3 exposed proteins (orange) and the genome (blue) is plotted. Wilcoxon two-sample test comparing  
4 sequence divergence of orthologs in the genome to the host-exposed proteins has p-value of  
5 6.8E-11.

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

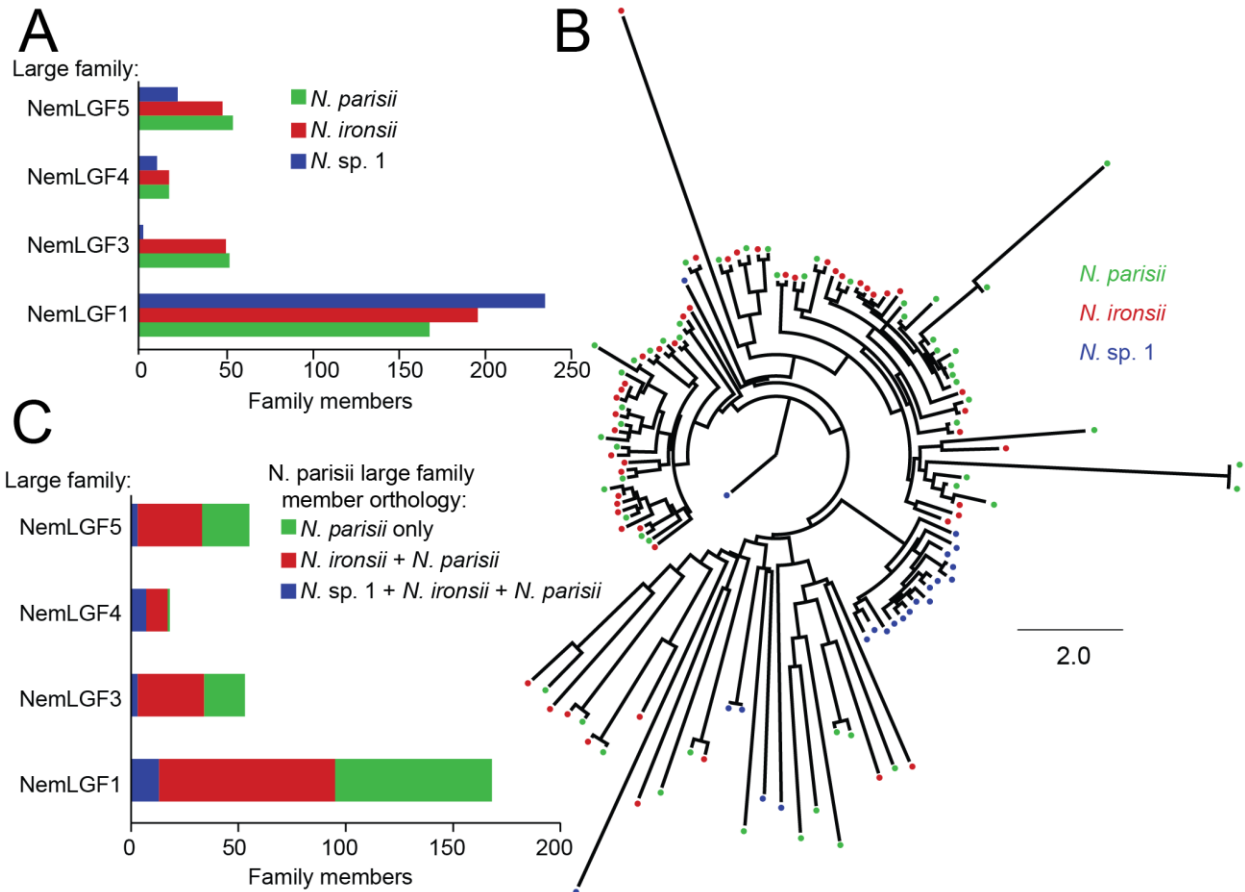
22

23

24

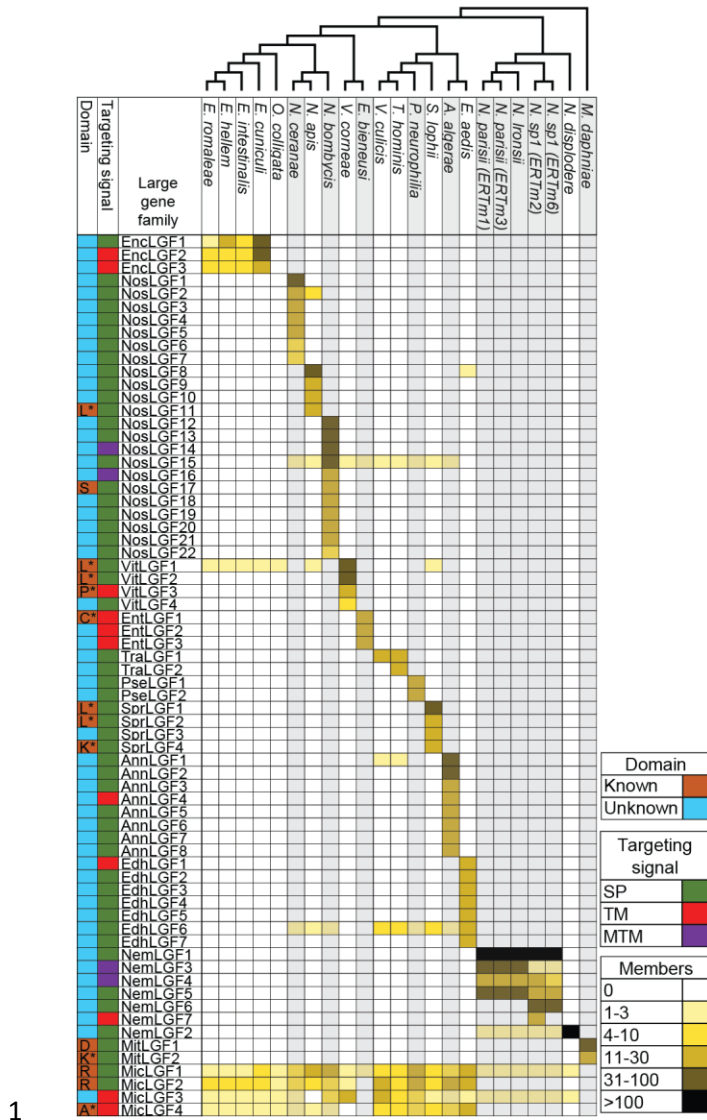
25

26



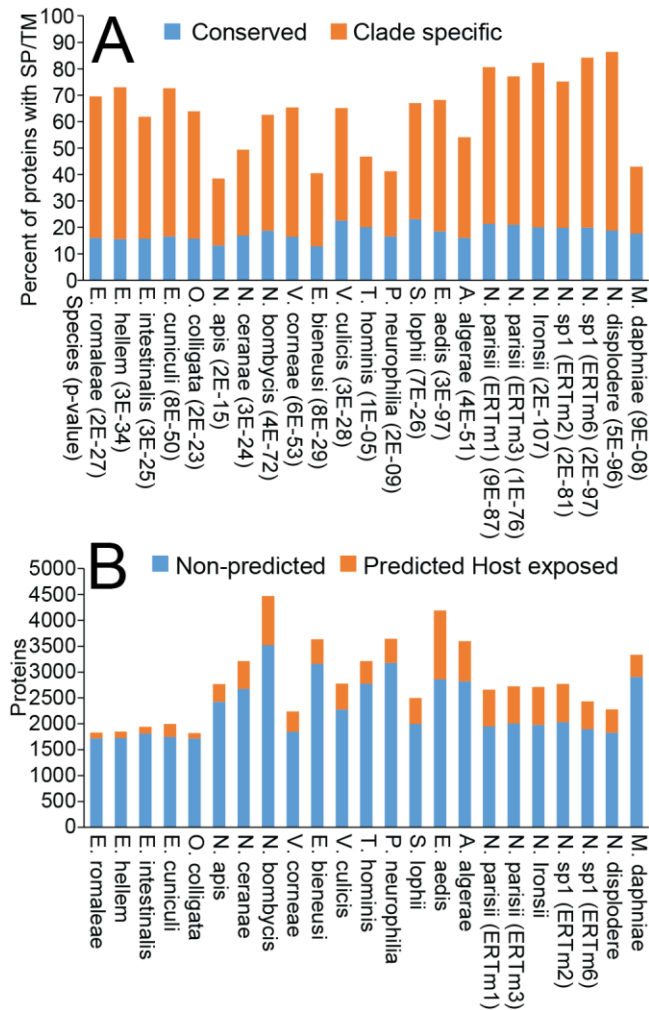
1  
2 **Figure 3. Species-specific radiation of large gene family members.** **A.** Number of members  
3 for each gene family that are present in each species. **B.** NemLGF5 tree showing gene family-  
4 specific radiation. Scale is changes per site. **C.** The number of proteins for the indicated gene  
5 family that are either unique to *N. parisii* (green), have orthologs in *N. ironsii* (red), or have  
6 orthologs in *N. sp. 1* and *N. ironsii* (blue).

7  
8  
9  
10  
11  
12



1  
2 **Figure 4. Large gene families are widespread throughout microsporidia.** Heat map showing  
3 large gene families identified in microsporidia and the number of gene family members in each  
4 species. Cladogram of species is shown at the top. Each column represents a species and strains  
5 are shown in parentheses. Each clade of species is alternatively shaded in grey or white. Each  
6 row represents a large gene family. Families are named and clustered based on the genus from  
7 where they were identified. The first column indicates if a known Pfam domain can be found within  
8 the indicated large gene family. Domains defined as follows: L (LRR), S (serpin), P (peptidase  
9 M48), C (chitin synthase), K (kinase), D (Duf3638), R (RicinB), and A (ABC transporter). Members  
10 of each gene family were determined using HMMER, except for those indicated with an \* which  
11 were determined using OrthoMCL. The second column indicates the targeting signal that is  
12 overrepresented within the indicated gene family. SP, signal peptide, TM, single transmembrane  
13 domain, and MTM, multitransmembrane domain. Each box in columns to the right of the gene  
14 family name is colored according to the total number of members within a given gene family.

15



1  
2 **Figure 5. Prediction of host-exposed microsporidia proteins.** **A.** Clade-specific microsporidia  
3 proteins (orange) are enriched in signal peptides/transmembrane domains compared to  
4 conserved proteins (blue). Enrichment p-values (one-sided Fisher's exact test) are listed in  
5 parenthesis below each species. **B.** The number of proteins in each microsporidia genome that  
6 are predicted to be host-exposed proteins (orange), compared to the rest of the genome (blue).

7  
8  
9  
10  
11  
12  
13  
14

## 1   **References**

- 2   1. Dean, P. Functional domains and motifs of bacterial type III effector proteins and their roles in  
3   infection. *FEMS Microbiol. Rev.* **35**, 1100–1125 (2011).
- 4   2. Blader, I. J. & Saeij, J. P. Communication between *Toxoplasma gondii* and its host: impact on  
5   parasite growth, development, immune evasion, and virulence. *APMIS Acta Pathol.*  
6   *Microbiol. Immunol. Scand.* **117**, 458–476 (2009).
- 7   3. van Ooij, C. *et al.* The malaria secretome: from algorithms to essential function in blood stage  
8   infection. *PLoS Pathog.* **4**, e1000084 (2008).
- 9   4. Aliberti, J. *et al.* Molecular mimicry of a CCR5 binding-domain in the microbial activation of  
10   dendritic cells. *Nat. Immunol.* **4**, 485–490 (2003).
- 11   5. Bougdour, A. *et al.* Host cell subversion by *Toxoplasma* GRA16, an exported dense granule  
12   protein that targets the host cell nucleus and alters gene expression. *Cell Host Microbe* **13**,  
13   489–500 (2013).
- 14   6. Saeij, J. P. J. *et al.* Polymorphic secreted kinases are key virulence factors in toxoplasmosis.  
15   *Science* **314**, 1780–1783 (2006).
- 16   7. Saeij, J. P. J., Arrizabalaga, G. & Boothroyd, J. C. A cluster of four surface antigen genes  
17   specifically expressed in bradyzoites, SAG2CDXY, plays an important role in *Toxoplasma*  
18   *gondii* persistence. *Infect. Immun.* **76**, 2402–2410 (2008).
- 19   8. Elsheikha, H. M. & Zhao, X. Patterns and role of diversifying selection in the evolution of  
20   *Toxoplasma gondii* SAG5 locus. *Parasitol. Res.* **103**, 201–207 (2008).
- 21   9. Rohmer, L., Guttman, D. S. & Dangl, J. L. Diverse evolutionary mechanisms shape the type  
22   III effector virulence factor repertoire in the plant pathogen *Pseudomonas syringae*. *Genetics*  
23   **167**, 1341–1360 (2004).
- 24   10. Singh, M. *et al.* Proteome analysis of *Plasmodium falciparum* extracellular secretory  
25   antigens at asexual blood stages reveals a cohort of proteins with possible roles in immune  
26   modulation and signaling. *Mol. Cell. Proteomics MCP* **8**, 2102–2118 (2009).

- 1 11. Malen, H., Berven, F. S., Fladmark, K. E. & Wiker, H. G. Comprehensive analysis of  
2 exported proteins from *Mycobacterium tuberculosis* H37Rv. *Proteomics* **7**, 1702–1718  
3 (2007).
- 4 12. Mahdavi, A. *et al.* Identification of secreted bacterial proteins by noncanonical amino  
5 acid tagging. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 433–438 (2014).
- 6 13. Rhee, H.-W. *et al.* Proteomic mapping of mitochondria in living cells via spatially  
7 restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013).
- 8 14. Vavra, J. & Lukes, J. Microsporidia and ‘the art of living together’. *Adv. Parasitol.* **82**,  
9 253–319 (2013).
- 10 15. Shaw, R. W., Kent, M. L. & Adamson, M. L. Innate susceptibility differences in chinook  
11 salmon *Oncorhynchus tshawytscha* to *Loma salmonae* (Microsporidia). *Dis. Aquat. Organ.*  
12 **43**, 49–53 (2000).
- 13 16. Pombert, J.-F., Haag, K. L., Beidas, S., Ebert, D. & Keeling, P. J. The *Ordospora*  
14 *colligata* genome: Evolution of extreme reduction in microsporidia and host-to-parasite  
15 horizontal gene transfer. *mBio* **6**, (2015).
- 16 17. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite  
17 *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
- 18 18. Stentiford, G. D., Feist, S. W., Stone, D. M., Bateman, K. S. & Dunn, A. M.  
19 Microsporidia: diverse, dynamic, and emergent pathogens in aquatic systems. *Trends*  
20 *Parasitol.* **29**, 567–578 (2013).
- 21 19. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating  
22 signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
- 23 20. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane  
24 protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*  
25 **305**, 567–580 (2001).

- 1 21. Desjardins, C. A. *et al.* Contrasting host-pathogen interactions and genome evolution in  
2 two generalist and specialist microsporidian pathogens of mosquitoes. *Nat. Commun.* **6**,  
3 7121 (2015).
- 4 22. Cuomo, C. A. *et al.* Microsporidian genome analysis reveals evolutionary strategies for  
5 obligate intracellular growth. *Genome Res.* **22**, 2478–2488 (2012).
- 6 23. Senderskiy, I. V., Timofeev, S. A., Seliverstova, E. V., Pavlova, O. A. & Dolgikh, V. V.  
7 Secretion of Antonospora (Paranosema) locustae proteins into infected cells suggests an  
8 active role of microsporidia in the control of host programs and metabolic processes. *PLoS*  
9 *One* **9**, e93585 (2014).
- 10 24. Campbell, S. E. *et al.* The genome of Spraguea lophii and the basis of host-  
11 microsporidian interactions. *PLoS Genet.* **9**, e1003676 (2013).
- 12 25. Troemel, E. R., Felix, M.-A., Whiteman, N. K., Barriere, A. & Ausubel, F. M.  
13 Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*.  
14 *PLoS Biol.* **6**, 2736–2752 (2008).
- 15 26. Luallen, R. J. *et al.* Discovery of a Natural Microsporidian Pathogen with a Broad Tissue  
16 Tropism in *Caenorhabditis elegans*. *bioRxiv* (2016). doi:10.1101/047720
- 17 27. Szumowski, S. C., Botts, M. R., Popovich, J. J., Smelkinson, M. G. & Troemel, E. R. The  
18 small GTPase RAB-11 directs polarized exocytosis of the intracellular pathogen *N. parisii* for  
19 fecal-oral transmission from *C. elegans*. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8215–8220  
20 (2014).
- 21 28. Estes, K. A., Szumowski, S. C. & Troemel, E. R. Non-lytic, actin-based exit of  
22 intracellular parasites from *C. elegans* intestinal cells. *PLoS Pathog.* **7**, e1002227 (2011).
- 23 29. Reid, A. J. Large, rapidly evolving gene families are at the forefront of host-parasite  
24 interactions in Apicomplexa. *Parasitology* **142 Suppl 1**, S57–70 (2015).



- 1 30. Balla, K. M., Andersen, E. C., Kruglyak, L. & Troemel, E. R. A wild *C. elegans* strain has  
2 enhanced epithelial immunity to a natural microsporidian parasite. *PLoS Pathog.* **11**,  
3 e1004583 (2015).
- 4 31. Spencer, W. C. *et al.* A spatial and temporal map of *C. elegans* gene expression.  
5 *Genome Res.* **21**, 325–341 (2011).
- 6 32. Nakjang, S. *et al.* Reduction and expansion in microsporidian genome evolution: new  
7 insights from comparative genomics. *Genome Biol. Evol.* **5**, 2285–2303 (2013).
- 8 33. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.  
9 *Nucleic Acids Res.* **44**, D279–285 (2016).
- 10 34. Reinke, A. W. & Troemel, E. R. The Development of Genetic Modification Techniques in  
11 Intracellular Parasites and Potential Applications to Microsporidia. *PLoS Pathog.* **11**,  
12 e1005283 (2015).
- 13 35. Hiller, N. L. *et al.* A host-targeting signal in virulence proteins reveals a secretome in  
14 malarial infection. *Science* **306**, 1934–1937 (2004).
- 15 36. Pan, G. *et al.* Comparative genomics of parasitic silkworm microsporidia reveal an  
16 association between genome expansion and host adaptation. *BMC Genomics* **14**, 186  
17 (2013).
- 18 37. Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E. S. & Keeling, P. J. The complete  
19 sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon*  
20 *intestinalis*. *Nat. Commun.* **1**, 77 (2010).
- 21 38. Solter, L. F. in *Microsporidia* 165–194 (John Wiley & Sons, Inc., 2014).
- 22 39. Lange, C. E. The host and geographical range of the grasshopper pathogen  
23 *Paranosema* (*Nosema*) *locustae* revisited. *J. Orthoptera Res.* **14**, 137–141 (2005).
- 24 40. Baltrus, D. A. *et al.* Dynamic evolution of pathogenicity revealed by sequencing and  
25 comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathog.* **7**, e1002132  
26 (2011).

- 1 41. Doyle, E. L., Stoddard, B. L., Voytas, D. F. & Bogdanove, A. J. TAL effectors: highly  
2 adaptable phytobacterial virulence factors and readily engineered DNA-targeting proteins.  
3 *Trends Cell Biol.* **23**, 390–398 (2013).
- 4 42. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system:  
5 biology, mechanisms and applications. *Biochimie* **117**, 119–128 (2015).
- 6 43. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing  
7 oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
- 8 44. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred  
9 kilobases. *Nat. Methods* **6**, 343–345 (2009).
- 10 45. Frokjaer-Jensen, C., Davis, M. W., Ailion, M. & Jorgensen, E. M. Improved Mos1-  
11 mediated transgenesis in *C. elegans*. *Nat. Methods* **9**, 117–118 (2012).
- 12 46. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans*  
13 genome using RNAi. *Nature* **421**, 231–237 (2003).
- 14 47. Reinke, A. W., Bennett, Eric & Troemel, E. R. Tissue and subcellular specific localization  
15 of proteins in *C. elegans* using spatially restricted enzymatic tagging. *Prep.*
- 16 48. Choi, H., Kim, S., Fermin, D., Tsou, C.-C. & Nesvizhskii, A. I. QPROT: Statistical method  
17 for testing differential expression using protein-level intensity data in label-free quantitative  
18 proteomics. *J. Proteomics* **129**, 121–126 (2015).
- 19 49. Bakowski, M. A. *et al.* Ubiquitin-mediated response to microsporidia and virus infection  
20 in *C. elegans*. *PLoS Pathog.* **10**, e1004200 (2014).
- 21 50. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.*  
22 **5**, R12 (2004).
- 23 51. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the  
24 genomic era. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361**, 1929–1940 (2006).
- 25 52. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for  
26 eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

- 1 53. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high  
2 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- 3 54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of  
4 large phylogenies. *Bioinforma. Oxf. Engl.* **30**, 1312–1313 (2014).
- 5 55. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function  
6 analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
- 7
- 8
- 9
- 10
- 11