Identification of microsporidia host-exposed proteins reveals a repertoire of rapidly evolving proteins

Aaron W. Reinke1, Keir M. Balla1, Eric J. Bennett1 & Emily R. Troemel1

Pathogens use a variety of secreted and surface proteins to interact with and manipulate their hosts, but a systematic approach for identifying such proteins has been lacking. To identify these ‘host-exposed’ proteins, we used spatially restricted enzymatic tagging followed by mass spectrometry analysis of Caenorhabditis elegans infected with two species of Nematocida microsporidia. We identified 82 microsporidia proteins inside of intestinal cells, including several pathogen proteins in the nucleus. These microsporidia proteins are enriched in targeting signals, are rapidly evolving and belong to large Nematocida-specific gene families. We also find that large, species-specific families are common throughout microsporidia species. Our data suggest that the use of a large number of rapidly evolving species-specific proteins represents a common strategy for microsporidia to interact with their hosts. The unbiased method described here for identifying potential pathogen effectors represents a powerful approach to study a broad range of pathogens.
Pathogens exploit hosts to promote their own proliferation. Viral, bacterial and eukaryotic pathogens control their hosts using effector proteins that interact directly with host molecules\(^1\)\(^-\)\(^3\). These effector proteins can be exported out of the pathogen into host cells or they can remain attached to the pathogen but with regions of the protein exposed to the host environment. These host-exposed proteins perform molecular functions that range from manipulation of host defenses to modulation of host pathways that can promote pathogen growth\(^3\)\(^-\)\(^5\). In many cases these proteins are evolving under diversifying selection, such that variation among these proteins can influence host survival\(^6\)\(^-\)\(^8\). Examples to date indicate considerable variation in the proteins that pathogens use to interface with their hosts. The conservation of these host-exposed proteins varies among different types of pathogens. Whereas most effectors of a strain of *Pseudomonas syringae* are present in other *Pseudomonas* strains and over 35% are conserved in other bacterial genera\(^9\), fewer than 15% of predicted host-exposed proteins of *Plasmodium falciparum* are reported to be conserved among *Plasmodium* species\(^10\).

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

Microsporidia constitute a large phylum of fungal-related obligate intracellular eukaryotic pathogens. The phylum contains over 1,400 described species that infect diverse animals including nematodes, arthropods and vertebrates, although individual species often have a narrow host range\(^14\)\(^-\)\(^15\). Dependent on their hosts for survival and reproduction, they have reduced genomes that lack several key regulatory and metabolic pathways\(^16\)\(^-\)\(^17\). Altogether these properties make microsporidia an excellent model of pathogen evolution. Despite the fact that microsporidia are of both medical and agricultural importance, tools for genetic modification of microsporidia are lacking and almost nothing is known about the proteins that enable interactions with their hosts\(^18\).

Two potential targeting signals are known that could expose microsporidia proteins to the host. These are N-terminal signal-sequences that direct proteins for secretion\(^19\), and transmembrane domains that could be used to attach proteins to the pathogen plasma membrane with regions of the microsporidia protein in direct contact with host molecules\(^20\). A number of studies have used these two targeting signals to predict the set of proteins encoded by pathogen genomes that are likely to be host-exposed\(^21\)\(^-\)\(^22\). However, it is unclear how accurate these approaches are at identifying such proteins in microsporidia and these prediction methods do not distinguish between proteins partially or wholly outside the microsporidia cell from those directed to internal membranes or compartments\(^13\). Although some host-exposed microsporidia proteins have been characterized\(^23\)\(^-\)\(^29\), no comprehensive identification of such proteins has been carried that include the intracellular stage of the pathogen.

Several microsporidia of the genus *Nematocida* naturally infect *C. elegans*, a model organism that offers a number of advantages for the study of host–pathogen interactions\(^30\)\(^-\)\(^31\). Infection of *C. elegans* by *N. parisi* begins with spores being ingested and then invading host intestinal cells. *N. parisi* initially develops in direct contact with the cytoplasm as a meront, eventually differentiating into a transmissible spore form that exits the cell\(^32\). Although the infection reduces worm lifespan, infected animals can generate enormous numbers of spores before death, with a single worm able to produce over 100,000 spores during the course of the infection\(^30\)\(^,\)\(^33\). Using *C. elegans*, we now report the first unbiased identification of microsporidia host-exposed proteins inside of an animal. These proteins are enriched for signal sequences and transmembrane domains, and they are rapidly evolving and tend to belong to unique large gene families. We also find that these species-specific large families are common throughout microsporidia. Using the properties we identified for the host-exposed proteins in *Nematocida*, we analysed 23 microsporidia genomes to predict potential host-exposed proteins, almost all were found to have no known molecular function. These results suggest that microsporidia use a set of lineage-specific, rapidly evolving proteins to interact with their hosts. This study provides a foundation for further functional characterization of host-exposed microsporidian proteins, and demonstrates the utility of proximity-labelling proteomic methods to broadly identify pathogen proteins localized within host cells.

Results

Identification of *Nematocida* host-exposed proteins. To identify microsporidia proteins that come into contact with the intracellular host environment, we used the technique of spatially restricted enzymatic tagging\(^13\). This approach uses the enzyme APX to label proteins in the compartment where the enzyme is expressed with a biotin handle for subsequent purification (Fig. 1a). We generated strains of *C. elegans* expressing GFP-APX, either in the cytoplasm or in the nucleus of intestinal cells (Fig. 1b). We also generated a negative control strain that expresses GFP in the intestine, but without the APX protein (Supplementary Table 1).

First, we inoculated these transgenic animals with *N. parisi* spores, which led to the majority of animals being infected (Supplementary Fig. 1). These animals were then incubated for 44 h at 20°C to allow for growth of the parasite. Next, we added the biotin-phenol substrate and hydrogen peroxide to these animals to facilitate APX-mediated biotinylation of host and pathogen proteins proximal to the GFP-APX protein. Under these conditions, we detected biotin-labelled proteins by microscopy in the intestinal cells of infected animals, but no labelling in the microsporidia cells themselves, demonstrating that the labelling technique is restricted to host-cell regions (Supplementary Fig. 2). Biotinylated proteins were isolated from total worm extracts using streptavidin-conjugated resin, and these purified proteins were identified using mass spectrometry. Biotinylated proteins from infected animals were isolated in triplicate and over 4,000 proteins from *C. elegans* and *N. parisi* were identified (Supplementary Fig. 3).

As validation that proteins were labelled in specific compartments in this experiment, we used the labelled *C. elegans* proteins as an internal control. By comparing spectral counts
If the host-exposed proteins we identified were truly secreted from *Nematocida* parasite cells into *C. elegans* host cells, we would predict that they would be processed by a signal peptidase in the parasite that would cleave off the signal sequence, and we could then detect the resulting N-terminal fragments in the host. Indeed, we detected N-terminal peptides corresponding to the predicted signal peptidase processed form for 4 of the 22 identified host-exposed proteins with signal peptides (two from *N. parisiis* and two from *N. sp. 1*—see below), providing support that microsporidia host-exposed proteins containing signal peptides are secreted into the host (Supplementary Data 2 and 3 and Supplementary Table 2).

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

**Many host-exposed proteins belong to large gene families.** Large, expanded gene families have been suggested to mediate host–pathogen interactions in a number of pathogen species and several large gene families have been previously identified in *Nematocida* species. We defined large gene families as groups of homologous proteins with at least 10 members in one species that were enriched in signal peptides or transmembrane domains. We initially identified these families from paralogous orthogroups and then generated profile hidden Markov models to identify additional members in the genome.

There are four large *N. parisiis* gene families that contain from 18 to 169 members. Two of these gene families, Nlmgf1 and Nlmgf5, encode signal peptides, and the other two gene families, Nlmgf3 and Nlmgf4, encode C-terminal transmembrane domains. The host-exposed proteins we identified are significantly enriched (P-value of 1.3E–16) in these families and contain 35 members of these four gene families, with at least one host-exposed protein in each of the four families (Fig. 2b,c). The four nuclear specific proteins are members of the Nlmgf1 or Nlmgf5 gene family, whereas four of the cytoplasmic specific proteins with transmembrane domains belong to the Nlmgf3 family (Supplementary Data 2).

*Nematocida* host-exposed proteins are clade-specific. To investigate how the repertoire of *N. parisiis* host-exposed proteins is evolving, we explored whether the identified host-exposed proteins are conserved in three other *Nematocida* species. The earliest known diverging species of the genus is *N. displodere*, which proliferates well in the epidermis and muscle, but poorly in the intestine. In contrast, the other *Nematocida* species are intestinal-specific. Previously, the species known to be the most closely related to *N. parisiis* was the intestinal-specific *N. sp. 1* strain ERTm2, which shares 68.3% average amino acid identity with *N. parisiis*. To provide a more closely related species for comparison, we sequenced and assembled the genome of *Nematocida* strain ERTm5, an intestinal-specific strain that was isolated from a wild-caught *C. briggsae* in Hawaii. This strain was previously described as a strain of *N. parisiis* based on rRNA

---

**Figure 1 | Overview of approach to detect and analyse host-exposed microsporidia proteins.** (a) Schematic of spatially restricted enzymatic tagging in *C. elegans*. Left, worms expressing GFP-APX in the cytoplasm of the intestine and infected with microsporidia are treated with biotin-phenol and H$_2$O$_2$. This treatment results in proteins within the intestinal cytoplasm being labelled with biotin. Right, an intestinal cell infected with microsporidia expressing cytoplasmic APX (green circular sectors) labelling microsporidia host-exposed proteins with biotin (red B). (b) Animals expressing GFP-APX in the intestine localized to either the cytoplasm (top) or the nucleus (bottom).
sequence, but based on our analysis, it now appears to define a new species (see ‘Methods’ section). This genome is comparable in quality to other sequenced genomes as judged both by assembly statistics and the presence of proteins conserved throughout microsporidia (Supplementary Table 3). This new species, *Nematocida ironsii*, now represents the closest known sister species to *N. parisii* and has an average amino acid identity of 84.7% compared with *N. parisii* (Supplementary Fig. 6 and Supplementary Table 5). To examine conservation, each *N. parisii* protein was placed into an orthogroup using six eukaryotic and 23 microsporidian genomes. Every *N. parisii* protein was categorized into one of six classes of decreasing conservation: (1) *N. parisii* proteins conserved with other non-microsporidia eukaryotes, (2) conserved with other eukaryotes, (3) conserved within *N. parisii*, (4) conserved within *N. displodere* and *N. sp. 1*, (5) conserved within *N. ironsii* and *N. displodere*, and (6) unique to *N. ironsii*. (a) Comparison of *N. parisii* proteins conserved with other non-microsporidia eukaryotes, (b) Comparison of *N. parisii* proteins conserved with other eukaryotes, (c) Comparison of *N. parisii* proteins conserved within the genome, (d) Comparison of *N. parisii* proteins conserved within *N. displodere* and *N. sp. 1*, (e) Comparison of *N. parisii* proteins conserved within *N. ironsii* and *N. displodere*, and (f) Comparison of *N. parisii* proteins unique to *N. ironsii*. The percentage of the genome and host-exposed proteins in each category is shown. TM, transmembrane; SP, signal peptide.
microsporidia, (3) conserved with *N. displodere*, (4) conserved with *N. sp. 1*, (5) conserved with *N. ironsii* and 6) those that are unique to *N. parisi* (Fig. 2d).

Using this evolutionary approach, we found that the set of host-exposed proteins we identified are significantly enriched (P-value of 1.9E – 20) for less conserved proteins, with only 12% having orthologs outside of a group of closely related *Nematocida* species (*N. sp. 1*, *N. ironsii* and *N. parisi*), which we refer to as ‘clade-specific’). In contrast, 63% of all *N. parisi* proteins in the genome have orthologs outside of this clade of *Nematocida* species (Fig. 2e). Most of these identified proteins don’t have a predicated molecular function, with only five of these 72 proteins containing a predicted Pfam domain (Fig. 2b). To determine the rate of protein evolution, we calculated the protein sequence divergence between orthologous *N. parisi* and *N. ironsii* proteins. We found that the host-exposed proteins are rapidly evolving compared with the other proteins in the genome (Fig. 2f).

To examine whether the properties of the host-exposed proteins we identified were conserved in other microsporidia species, we performed spatially restricted enzymatic tagging on *C. elegans* infected with *N. sp. 1*. Although we identified fewer *C. elegans* and microsporidia proteins from *N. sp. 1* infected animals, we nonetheless found ten proteins enriched over background (Supplementary Fig. 3 and Supplementary Data 3). These proteins have similar properties to those identified for *N. parisi* and *N. ironsii* proteins. We next investigated whether proteins that are not widely conserved in *N. parisi* as they are enriched in targeting signals and clade-specific proteins (that is, proteins not conserved in other eukaryotes, microsporidia or *N. displodere*) (Supplementary Fig. 7). They also are enriched for being members of large gene families, including three members of NemLGFI and one member of the *N. sp. 1*-specific family NemLGF6. We also identified two pairs of orthologs from the two species: hexokinase (NEPG_02043 and NERG_02003) and a NemLGFI family member (NEPG_02370 and NERG_01049). To expand this analysis to a different microsporidia genus, we examined data previously generated from germinated *Spraguea lophii* spores. We found that proteins identified as secreted from these germinated spores were also enriched in the properties of signal peptides and clade-specific proteins (Supplementary Fig. 8)29.

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

Current limitations of proteomic methods suggest that this approach will not result in the complete identification of all host-exposed microsporidia proteins. To estimate the sensitivity of this method we compared the identified *C. elegans* intestinal proteins to the total number of mRNAs expressed in the intestine38. We also compared the total number of detected *N. parisi* proteins to the number encoded by the proteome. From these comparisons, we estimate that we identified between ~8 and 24% of potential host-exposed proteins. This would mean that the total host-exposed proteome encoded by *N. parisi* is in the order of 300–900 proteins, a range that encompasses the number of proteins in the genome that have the properties enriched in the experimentally identified host-exposed proteins.

**Large families display lineage-specific expansions.** If most members of *N. parisi* large gene families are involved in host–pathogen interactions, we would predict that they would also be rapidly evolving with species-specific radiations. The four large gene families of *N. parisi* contain a total of 295 members. Members of these four families are also present in the other species in this clade, *N. sp. 1* and *N. ironsii*, but not any other microsporidia species (Figs 3a and 4). Phylogenetic trees of these families show expansion of family members specific to each species (Fig. 3b,c). Members from these families are often not conserved between species, with only 5–39% of *N. parisi* members in each gene family that have orthologs in *N. sp. 1* and 56–95% that have orthologs in *N. ironsii* (Fig. 3e). The largest families that have signal peptides, NemLGFI and NemLGFS, are enriched for genes on the ends of chromosomes, a chromosomal localization that is not enriched in the transmembrane-containing families (Supplementary Fig. 9A). The four families are often adjacent to each other, suggesting they are being generated through local duplication events (Supplementary Fig. 9B).

**Large families are common in microsporidia.** To examine whether large gene families are common in other microsporidia species, we examined 23 microsporidia genomes (17 other microsporidia species and six from *Nematocida*) (Supplementary Fig. 6). From these 21 species, 68 families were identified with at least 10 members in one species and enriched in either predicted signal peptides or transmembrane domains. In addition, we found that most (59 of 68) of these families do not have any members present outside of the genus or species. For example, there are three families with members present in all four *Encephalitozoon* species but no other species examined. In addition, we identified four large gene families that were conserved throughout most microsporidia including two ricin B domain-containing families29. All but one species examined has a large genus-specific family, demonstrating that large gene families are widespread throughout microsporidia.

**Prediction of host-exposed proteins from other microsporidia.** We next investigated whether proteins that are not widely conserved in microsporidia share properties with the identified host-exposed proteins. We examined 23 microsporidian genomes to identify proteins that are not conserved with other eukaryotes, or conserved with distantly related microsporidia species. These clade-specific proteins are all significantly enriched in targeting signals compared with proteins conserved with more distally related microsporidia or other eukaryotes (Fig. 5a). This result is similar to what we found in our analysis of experimentally identified host-exposed proteins in *Nematocida*, and similar to a previous study of several microsporidian species37.
putative host-exposed proteins do not have a predicted molecular function, with only 7.4% having a predicted Pfam domain that occurs in proteins outside of microsporidia (Supplementary Data 5). Although most of these proteins do not have known domains, several species have expanded families of leucine-rich repeat (LRR) domains and two species have expanded families of protein kinases (Fig. 5). The most frequently observed domains in putative host-exposed proteins that are not members of the large gene families are transporters, kinases, LRR domains, ubiquitin carboxyl-terminal hydrolases and the bacterial specific DUF1510 (Supplementary Fig. 10)38. Interestingly, a number of domains that are present in the large gene families are also observed in the non-paralogous proteins, suggesting that there are several common domains that have been utilized in multiple microsporidia species to interact with hosts. These predictions of host-exposed proteins suggest that microsporidia employ a large number of proteins with novel domains to interact with hosts.

**Discussion**

To understand how microsporidia interact with their hosts, we experimentally identified 82 host-exposed proteins from two *Nematocida* species. To identify these proteins, we employed an unbiased approach that labelled the host-exposed pathogen proteins, we found that four proteins were specific to the nucleus, proteins inside of an intact animal. Interestingly, of the identified proteins, we had been previously reported, here we provide a comprehensive identification of these gene families throughout microsporidia29,31,44,45. The majority of these large gene families have no known molecular function based on sequence similarity. One enticing possibility is that the expansion of these families is due to interactions with host proteins. In support of this possibility, a number of the gene families with predicted domains are known to mediate protein–protein interactions including LRR and RING domains.

One intriguing characteristic of these large gene families is that they are either genus- or species-specific, with large lineage-specific expansions of these gene families across microsporidia. The differences in the total number of gene families can be quite large in the same genus. For example, in the family NemLGF3, *N. sp. 1* only has three members compared with 53 members in *N. parisii*. Both strains of *N. sp. 1* also have a gene family
### Figure 4 | Large gene families are widespread throughout microsporidia.

Heat map showing large gene families identified in microsporidia and the number of gene family members in each species. Cladogram of species is shown at the top. Each column represents a species and strains are shown in parentheses. Each clade of species is alternatively shaded in grey or white. Each row represents a large gene family. Families are named and clustered based on the genus from which they were identified. The first column indicates if a known Pfam domain can be found within the indicated large gene family. Domains defined as follows: L (LRR), S (serpin), P (peptidase M48), C (chitin synthase), K (kinase), D (Duf3638), R (RicinB) and A (ABC transporter). Members of each gene family were determined using HMMER, except for those indicated with an *, which were determined using OrthoMCL. The second column indicates the targeting signal that is overrepresented within the indicated gene family. SP, signal peptide, TM, single transmembrane domain and MTM, multitransmembrane domain. Each box in columns to the right of the gene family name is coloured according to the total number of members within a given gene family.
microsporidia species there is both ecological evidence and laboratory studies demonstrating that the same strain of microsporidia can infect closely related host species\cite{3,46,47}. We speculate this host diversity could drive the expansion of large gene families in microsporidia and that these large gene families may in turn influence the host range.

The majority of the host-exposed proteins we identified in *N. parisi* and *N. sp*. 1 were proteins not conserved with *N. displodere* or other microsporidia species. Although lack of conservation accounts for most of the proteins identified, several conserved proteins were identified, including hexokinase, which we identified in both *Nematocida* species. Hexokinase was previously found to have predicted signal peptides in several microsporidia species and to be secreted in a heterologous system, and to be secreted from the microsporidia *Antonospora locustae*, providing experimental evidence that secreted hexokinase is a conserved feature of microsporidia\cite{22,28,37}. There are also several large gene families that have members present in multiple microsporidia species. This observation suggests that although selective forces result in a host-exposed protein repertoire with many unique proteins for each microsporidia clade, there are some proteins conserved throughout microsporidia involved in host interactions.

A number of forces are likely to shape the repertoire of host-exposed proteins, including the selective pressure of the host and interactions with other pathogens. Many of the features of the host-exposed protein repertoire in microsporidia are similar to characteristics reported in the apicomplexan phylum of protozoan obligate intracellular pathogens. Large gene families with either signal peptides or transmembrane domains are common. These families often have subtelomeric genomic locations and are species specific\cite{34}. Over 200 secreted proteins have been predicted in *P. falciparum* and few are conserved with other *Plasmodium* species\cite{5}. Most of these proteins also have no predicted molecular function\cite{43}. These similarities among species suggest that similar selective pressures can sculpt a host-exposed protein repertoire with related properties. In contrast, strains of the bacterium *P. syringae* are predicted to have less than 40 type III effectors, many of which are shared with other bacteria and display evidence of horizontal gene transfer\cite{9,48}.

A striking result of our analysis is that a large number of experimentally identified and predicted host-exposed proteins do not have domains found outside of microsporidia. These host-exposed proteins are a potential source of novel biochemical activity as the extreme selective pressures inflicted on pathogens by the host has been shown to result in unique molecular functions\cite{19,50}. Interestingly, we also predict a large per cent of the microsporidia genome to be responsible for mediating host–pathogen interactions. This suggests that although microsporidia have the smallest known genomes of any eukaryotes, they somewhat paradoxically encode a substantial cadre of proteins for interacting with their hosts. Understanding how microsporidia use these proteins to mediate host interactions will provide insight into their impact on hosts and the constraints on evolution of a minimalistic eukaryotic genome.

### Methods

**Cloning and generation of ***C. elegans*** expressing APX.*** Soybean APX (W41F) was optimized for *C. elegans* expression using DNAworks to design primers\cite{21}. These primers were annealed using a two-step PCR method and cloned into Gateway plasmid PDONR 221. Gibson cloning was then used to introduce GFP as an N-terminal fusion, and NES (LQLPPLERLTLD) and NLS (PKKKRKVD) tags to the C-terminus of APX\cite{22}. One kilobase (kb) of sequence upstream of the intestinal-specific gene *scp-5* was used as a promoter and *unc-54* as a 3’ sequence. Multisite Gateway was used to combine these fragments into the plasmid pCF150 to generate targeting constructs. The MosSCI approach was used to generate single copy insertions by injecting *unc-119* mutants from the EGI699 strain with these targeting constructs\cite{27,28}. Each transgenic strain was

![Figure 5](https://example.com)
Spatially restricted enzymatic tagging in C. elegans strains that express GFP-APX either localized to the cytoplasm or nucleus, as well as a control strain expressing GFP only, were used in subsequent experiments.

Peptide and protein identification and quantification. The resultant RAW files were converted into mzXML format using the ReadW.exe programme. The SEQUEST search algorithm (version 28) was used to search MS/MS spectra against an A concatenated target-decoy database comprised of forward and reverse sequences. The SEQUEST search algorithm (version 28) was used to search MS/MS spectra against a concatenated target-decoy database comprised of forward and reverse sequences. Identifications were filtered to a peptide false discovery rate of 2% using the linear model that is representative of individual analyses. Peptides were first separated by a concatenated target-decoy database comprised of forward and reverse sequences. The SEQUEST search algorithm (version 28) was used to search MS/MS spectra against a concatenated target-decoy database comprised of forward and reverse sequences. 

Analysis of mass spectrometry data. The peptide spectral counts of proteins were used to calculate fold change ratios and FDR P-values between GFP only, NES and NLS samples with the best model and a cutoff of 0.37 for both the noTM model and for the TM model. These same cutoffs were used to generate signal peptide processed forms.

Microscopy of infected C. elegans. To detect biotin labelling in infected worms, intestines were dissected and stained with anti-GFP (Roche) and Streptavidin Alexafluor 568 (Thermo Fisher). Images were taken using a Zeiss LSM700 confocal microscope with a x40 objective. To detect microsporidia in infected worms, fluorescence in situ hybridization with probes specific for microsporidia was performed as previously described and imaged with a Zeiss Axiosmager M1 microscope.

Microscopy of infected C. elegans. To detect biotin labelling in infected worms, intestines were dissected and stained with anti-GFP (Roche) and Streptavidin Alexafluor 568 (Thermo Fisher). Images were taken using a Zeiss LSM700 confocal microscope with a x40 objective. To detect microsporidia in infected worms, fluorescence in situ hybridization with probes specific for microsporidia was performed as previously described and imaged with a Zeiss Axiosmager M1 microscope.

Analysis of mass spectrometry data. The peptide spectral counts of proteins were used to calculate fold change ratios and FDR P-values between GFP only, NES and NLS samples with the best model and a cutoff of 0.37 for both the noTM model and for the TM model. These same cutoffs were used to generate signal peptide processed forms.

Microscopy of infected C. elegans. To detect biotin labelling in infected worms, intestines were dissected and stained with anti-GFP (Roche) and Streptavidin Alexafluor 568 (Thermo Fisher). Images were taken using a Zeiss LSM700 confocal microscope with a x40 objective. To detect microsporidia in infected worms, fluorescence in situ hybridization with probes specific for microsporidia was performed as previously described and imaged with a Zeiss Axiosmager M1 microscope.
Determination of conservation. For N. parisii proteins, conservation was determined based on orthogroups, except for the large gene families NemLGFI and NemLGFE-4 for which orthology was determined as described above. The following procedure was used to place the N. parisii proteins into six categories. If a N. parisii protein was in any group with a protein from the six non-microsporidian eukaryotic species, the protein was placed in the category ‘Eukaryotes’. If any remaining unassigned proteins were in a group with a protein from the microsporidia species not in the genus Nemacodida, then it was placed in the category ‘microsporida’. If any remaining unassigned proteins were in a group with an N. diplodore protein, then it was placed in the category ‘N. diplodore’. If any remaining unassigned proteins were in a group with an N. ironsii protein, then it was placed in the category ‘N. ironsii’. The remaining proteins were placed in the category ‘N. parisii’.

To predict host-exposed proteins the conservation of microsporidia proteins was determined. Proteins of each species were placed into two classes, ‘Conserved’ or ‘clade specific’. If a protein was in the same group as a protein from the eukaryotic or microsporidia species then it was classified as ‘conserved’. Otherwise it was classified as ‘clade-specific’. This was done except for the closer-related species, where proteins in the same clade were not considered. For this purpose the following clade definitions were used: Nematocidae species are N. parisii, N. sp. 1, and N. ironsii; Encephalitozoon species are E. romaleae, E. hellem, E. intestinalis, E. cuniculi and O. colligate; and the species V. culuis and T. hominis.

Calculation of protein sequence divergence. Proteins for microsporidia genomes were placed into orthogroups as described above. Proteins from one-to-one orthologs of the two N. parisii strains (ERTm1 and ERTm3) and N. ironsii were aligned using MUSCLE 3.8.31 (ref. 62). For large gene families orthology were determined as described above. For proteins conserved with N. sp. 1, the evolution rate was only calculated for one-to-one orthologs between the five genomes. For proteins conserved with N. diplodore, the evolution rate was only calculated for one-to-one orthologs between all six Nematocidae genomes. Maximum likelihood trees were built using ortholog sets (three sequences per set) of aligned protein sequences using PHYLIPI (http://evolution.genetics.washington.edu/phylip.html). The sum of the sequence tree length dived by the number of sequences in, PAM units, was calculated for each orthogroup set.

Data availability. The Whole Genome Shotgun project for the N. ironsii genome has been deposited at DDBJ/ENA/GenBank under the accession LTDK00000000. The version described in this paper is version LTDK01000000. All data supporting this manuscript is available from the corresponding author upon request.

References
1. Dean, P. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. FEMS Microbiol. Rev. 35, 1100–1125 (2011).
2. Blader, I. J. & Saeji, J. P. Communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion, and virulence. APMS Acta Pathol. Microbiol. Immunol. Scand. 117, 458–476 (2009).
3. van Ooij, C. et al. The malaria secretome: from algorithms to essential function in blood stage infection. PLoS Pathog. 4, e1000804 (2008).
4. Aliberti, J. et al. Molecular mimicry of a CCR5 binding-domain in the microbial activation of dendritic cells. Proc. Natl Acad. Sci. USA 109, 21106–21110 (2012).
5. Rohmer, L., Guttman, D. S. & Dangl, J. L. Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. Genetics 167, 1341–1360 (2004).
6. Sinha, M. et al. Proteomic analysis of Plasmodium falciparum extracellular secretory antigens at asexual blood stages reveals a cohort of proteins with possible roles in immune modulation and signaling. Mol. Cell Proteomics 8, 2102–2118 (2009).
39. Escoll, P., Mondino, S., Rolando, M. & Buchrieser, C. Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. Nat. Rev. Microbiol. 14, 5–19 (2016).
40. Reinek, A. W. & Troemel, E. R. The development of genetic modification techniques in intracellular parasites and potential applications to microsporidia. PLoS Pathog. 11, e1005283 (2015).
41. Chen, C.-L. et al. Proteomic mapping in live Drosocephila tissues using an engineered ascorbate peroxidase. Proc. Natl Acad. Sci. USA 112, 12093–12098 (2015).
42. Hung, V. et al. Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. Nat. Protoc. 11, 456–475 (2016).
43. Hiller, N. L. et al. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306, 1934–1937 (2004).
44. Pan, G. et al. Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. BMC Genomics 14, 186 (2013).
45. Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E. S. & Keeling, P. J. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitococcus intestinalis. Nat. Commun. 1, 77 (2010).
46. Solter, L. F. in Microsporidia 165–194 (John Wiley & Sons, Inc., 2014).
47. Lange, C. E. The host and geographical range of the grasshopper pathogen Pseudomonas (Nosema) locustae revisited. J. Orthoptera Res. 14, 137–141 (2005).
48. Baltrus, D. A. et al. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 7, e1002132 (2011).
49. Doyle, E. L., Stoddard, B. L., Voytas, D. F. & Bogdanove, A. J. TAL effectors: highly adaptable phytobacterial virulence factors and readily engineered DNA-targeting proteins. Trends Cell Biol. 23, 390–398 (2013).
50. Reth, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: biology, mechanisms and applications. Biochemie 117, 119–128 (2015).
51. Hoover, D. M. & Lukbowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res. 30, e43 (2002).
52. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 345–349 (2009).
53. Prokjaer-Jensen, C., Davis, M. W., Ailion, M. & Jorgensen, E. M. Improved Mos1-mediated transgenesis in C. elegans. Nat. Methods 9, 117–118 (2012).
54. Kamath, R. S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421, 231–237 (2003).
55. Reinek, A. W., Mak, R., Troemel, E. R. & Bennett, E. J. In vivo mapping of tissue- and subcellular-specific proteomes in Caenorhabditis elegans. Preprint at bioRxiv. http://dx.doi.org/10.1101/066134 (2016).
56. Huttlin, E. L. et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell. 143, 1174–1189 (2010).
57. Choi, H., Kim, S., Fermin, D., Tsonu, C.-C. & Nesvizhskii, A. I. QPROT: statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. J. Proteomics 129, 121–126 (2015).
58. Bakowski, M. A. et al. Ubiquitin-mediated response to microsporidia and virus infection in C. elegans. PLoS Pathog. 10, e1004200 (2014).
59. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
60. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361, 1929–1940 (2006).
61. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).
62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
63. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

Acknowledgements
We thank Steven Wasserman, Matthew Daugherty, Robert Luallen, Kirthi Reddy, Michael Botts, and Lianne Cohen for providing helpful comments on the manuscript. A.W.R. is a Monsanto Fellow of the Life Sciences Research Foundation. Some C. elegans strains were provided by the Caenorhabditis Genetics Center, which is funded by National Institutes of Health (NIH) Office of Research Infrastructure Programs Grant P40 OD610440. This work was supported by a New Scholar in Aging award from the Ellison Medical Foundation (E.J.B.) and R01GM114139, the David and Lucile Packard Foundation and a Burroughs Wellcome Fund fellowship (E.R.T.).

Author contributions
A.W.R. designed, conducted and analysed experiments. K.M.B. provided the N. ironsii genome sequence. E.J.B. performed the mass spectrometry analysis. E.R.T. provided mentorship and with A.W.R. and E.J.B. prepared the manuscript.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Reinek, A. W. et al. Identification of microsporidia host-exposed proteins reveals a repertoire of rapidly evolving proteins. Nat. Commun. 8, 14023 doi: 10.1038/ncomms14023 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.