EC Journals.ASM.org

Complete Genome Sequences from Three Genetically Distinct Strains Reveal High Intraspecies Genetic Diversity in the Microsporidian *Encephalitozoon cuniculi*

Jean-François Pombert,[a] Jinshan Xu,[a] David R. Smith,[a] David Heiman,[b] Sarah Young,[b] Christina A. Cuomo,[b] Louis M. Weiss,[c] Patrick J. Keeling[a]

Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada[a]; The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA[b]; Department of Pathology, Division of Parasitology, and Department of Medicine, Division of Infectious Diseases, Albert Einstein College of Medicine, Bronx, New York, USA[c]

Microsporidia from the Encephalitozoonidae are obligate intracellular parasites with highly conserved and compacted nuclear genomes: they have few introns, short intergenic regions, and almost identical gene complements and chromosome arrangements. Comparative genomics of *Encephalitozoon* and microsporidia in general have focused largely on the genomic diversity between different species, and we know very little about the levels of genetic diversity within species. Polymorphism studies with *Encephalitozoon* are so far restricted to a small number of genes, and a few genetically distinct strains have been identified; most notably, three genotypes (ECI, ECII, and ECIII) of the model species *E. cuniculi* have been identified based on variable repeats in the rRNA internal transcribed spacer (ITS). To determine if *E. cuniculi* genotypes are genetically distinct lineages across the entire genome and at the same time to examine the question of intraspecies genetic diversity in microsporidia in general, we sequenced *de novo* genomes from each of the three genotypes and analyzed patterns of single nucleotide polymorphisms (SNPs) and insertions/deletions across the genomes. Although the strains have almost identical gene contents, they harbor large numbers of SNPs, including numerous nonsynonymous changes, indicating massive intraspecies variation within the Encephalitozoonidae. Based on this diversity, we conclude that the recognized genotypes are genetically distinct and propose new molecular markers for microsporidian genotyping.

The nuclear genome of the microsporidian parasite *Encephalitozoon cuniculi* strain GB-M1 was the first to be characterized from any microsporidian, and at only 2.9 Mbp and roughly 2,000 genes (1), it has become a model for extreme reduction and the minimum genetic information that a pathogenic eukaryote needs to survive. This genome lacks metabolic pathways that were once thought to be essential for eukaryotes, and it has acquired, through horizontal transfer, genes encoding transporters that harness energy and metabolites from the host (2). Whole-genome sequencing has also revealed a high degree of streamlining in several other microsporidia, including congeners *E. hellem* (2.5 Mbp), *E. romaleae* (2.5 Mbp), and *E. intestinalis* (2.3 Mbp), the last of which has the smallest nuclear genome on record (3, 4). The differences in genome size among *Encephalitozoon* taxa are primarily due to variations in subtelomeric regions, and the four species have otherwise almost identical gene contents and chromosome arrangements. Their 11 chromosomes are extremely gene dense, with over 90% of their cores composed of coding loci and genes characterized by a paucity of introns.

Although comparative genomics has given us a good understanding of the genomic diversity among *Encephalitozoon* species, we know very little about the genetic/genomic diversity within species. Microsporidian polymorphism studies have focused largely on the human pathogen *Enterocytozoon bieneusi*, for which >80 different genotypes are known (see, e.g., references 5, 6, 7, and 8), and the honeybee and silkworm parasites from the genus *Nosema* (see, e.g., references 9, 10, and 11). *Enterocytozoon* and *Nosema* have more expanded genomes (6 to 10 Mbp) than *Encephalitozoon* species, implying different evolutionary constraints, such that their variability may not parallel that of their *Encephalitozoon* relatives. Moreover, only a few distinct *Encephalitozoon* genotypes have been described: 3 for *E. cuniculi*, 2 for *E. hellem*, and only 1 for both *E. intestinalis* and *E. romaleae* (12–17). None of these has been compared at the genome level; indeed, most studies on within-species diversity of microsporidia are limited to one or a few loci, such as the internal transcribed spacer (ITS) region between rRNA-coding genes. In fact, the only published genome-wide investigation of microsporidia involving closely related strains (18) focused on ploidy level and heterozygosity within strains and did not investigate polymorphisms between strains in detail.

Here, we examine the genetic diversity between complete genomes from three isolates of *E. cuniculi*, a zoonotic species infecting a wide range of mammals (19), to see how much genetic variability exists within the species. These isolates represent three distinct genotypes (ECI, ECII, and ECIII) developed for diagnostic purposes and defined by the number of GTTT repeats encoded within the ITS locus (14). From complete genome sequences we surveyed genome-wide levels and distribution of single nucleotide

Received 7 November 2012 **Accepted** 27 December 2012

Published ahead of print 4 January 2013

Address correspondence to Patrick J. Keeling, pkeeling@mail.ubc.ca.

J.-F.P. and J.X. contributed equally to this article.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/EC.00312-12.

doi:10.1128/EC.00312-12

polymorphisms (SNPs) and insertion-deletion events (indels). Overall, we find substantial interstrain diversity within *E. cuniculi* as well as remarkably high levels of interspecies diversity with the Encephalitozoonidae. SNPs are distributed more or less evenly across all chromosomes in the three genotypes, confirming that the variable repeats in ITS do represent the genome as a whole for three coherent genetically distinct populations. These analyses also suggest other potential molecular markers for microsporidian genotyping that may have greater resolution, and they raise some interesting questions regarding the architecture of the *E. cuniculi* genome, including the origin of G+C shifts, the location of centromeres, and the presence of sex-determining loci.

MATERIALS AND METHODS

Tissue culture and DNA purification. *Encephalitozoon cuniculi* genotypes ECI (rabbit isolate, ATCC 50503 [20]), ECII (mouse isolate [21]), and ECIII (canine isolate, ATCC 50502 [22]) were cultured in T25 or T75 flasks at 37°C and 5% CO_2 in RK13 cells (CCL37; American Type Culture Collection, Manassas, VA). Infected RK13 cells were maintained in continuous culture in minimum essential medium (MEM) (Invitrogen, Carlsbad, CA) supplemented with 7% heat-inactivated fetal calf serum and 1% penicillin-streptomycin-amphotericin B (Invitrogen, Carlsbad, CA). Cultures were subpassaged every 3 weeks. Supernatants from infected flasks containing microsporidian spores were collected twice weekly and replaced with fresh medium.

Supernatants containing spores were stored at 4°C until extraction of DNA. To enrich spores from host cell debris, the collected culture supernatants were subjected to sequential washes at 400 *g* each with distilled H_2O, Tris buffered saline (TBS)–Tween 20 (0.3%), and TBS. Spores were then filtered through a Nuclepore 3.0-μm filter to remove residual host cells. Spore concentrations were determined by counting spores using a hemocytometer (Improved Neubauer).

Spores were pelleted by centrifugation, resuspended in 300 μl of lysis solution (Epicentre, Madison, WI) containing proteinase K, and mixed thoroughly using a vortex mixer. Glass beads (200 μl, 150 to 212 μm in diameter) were added to the samples, which were immediately incubated at 65°C for 15 min and bead beaten at 2,500 rpm for 30 s every 5 min. The samples were then cooled to 37°C and incubated for 30 min at the same temperature upon the addition of RNase A (10 μg total). After treatment with RNase, the samples were placed on ice for 5 min, 150 μl of MPC protein precipitation reagent (Epicentre, Madison, WI) was added per sample, and the solutions were vortexed vigorously for 10 s. Protein debris was pelleted at 4°C for 10 min at ≥10,000 × *g*, and the supernatants were transferred to clean microcentrifuge tubes. DNA was then precipitated using isopropanol and rinsed twice using 70% ethanol, and the DNA was finally suspended in Tris-EDTA (TE) buffer.

Sequencing. DNAs from *E. cuniculi* ECI, ECII, and ECIII strains were sequenced using the Illumina HiSeq 2000 platform by the Broad Institute (101-bp paired ends; 220-bp inserts; average standard deviation, 75 to 79).

Genome assembly. For three strains of *E. cuniculi*, more than 1,000× sequence coverage was generated using 101b Illumina reads from 180b fragments (after error correction of the 220-bp inserts; see below). Sequences were filtered using multiple approaches to either remove contaminating sequence or identify reads matching the source organism. A scan of a small subset of reads and draft assemblies by BLAST (23) against the NCBI nonredundant nucleotide database indicated contamination from the host cell line (RK13) and bacteria; the raw data were then filtered by aligning with BWA (24) to the reference sequences for these contaminating genomes.

The remaining reads were initially assembled using a protocol established for other microsporidian genomes sequenced at the Broad Institute. Reads were first processed by ALLPATHS-LG (25) to create error-corrected 180b filled fragments. Reads were also filtered by selecting those with BWA alignment hits matching *E. cuniculi* draft assemblies that we generated using different assembly algorithms and the previously published GB-M1 reference genome. Using the selected filled fragment reads, a new assembly was generated with Newbler runAssembly (Roche, Branford, CT), which was tested against the unpaired 180b filled fragments generated by ALLPATHS and found to be the best-performing assembly algorithm evaluated. This was merged using Minimus2 from the AMOS package (26) with the best contamination-filtered assembly (Velvet [27] or ABySS [28]) of the raw reads to capture missing sequence regions, the consensus was corrected using Newbler runMapping with the error-corrected filled fragments, and a last check for missing regions was evaluated by comparing this assembly to the Newbler runAssembly to improve overall statistics, including total sequence. The resulting assemblies were evaluated by aligning using Nucmer from the MUMmer3 package (29) to the GB-M1 reference to confirm the absence of possible misassemblies or rearrangements. All resulting contigs in the final version had a BLAST match to GB-M1.

These initial drafts contained many apparent breaks in the 11 chromosomes, so further assembly and polishing steps were carried out following protocols established for other *de novo* Illumina-sequenced *Encephalitozoon* species (3, 4). First, paired-end reads from each of the three *E. cuniculi* strains were assembled *de novo* with Ray 1.6.1 rc2 (30) using iterative k-mer values of 21 to 31 on 8 processing cores (2 Intel Xeon E5506 CPUs at 2.13 GHz) with a maximum RAM allowance of 96 Gb. The resulting contigs were filtered by size with sort_contigs.pl (Advanced Center for Genome Technology, University of Oklahoma [www.genome.ou.edu/informatics.html]), and contigs equal to or longer than 500 bp were used as canvas to generate a BLAST (23) database with MAKEBLASTDB from the NCBI BLAST 2.2.26 package. Contigs constituting the cores of the chromosomes were identified by BLAST homology searches using the *E. cuniculi* strain GB-M1 genome as query, pulled out from the multifasta assembly file with the command line utility faSomeRecords (UCSC Genome Bioinformatics, University of California, Santa Cruz [http://genome.ucsc.edu/]), and concatenated into a single file (one for each ECI, ECII, and ECIII strain).

The draft assemblies and the new paired-end contigs were compared and merged with CONSED 22 (31). Subsets of the Illumina reads were iteratively mapped back on the merged contigs using Sanger quality scores with the addSolexaReads.pl script from the CONSED package, modified to increase the mapping stringency (i.e., −minmatch, 50; −minscore, 50; −penalty, −9). Contigs were extended according to the paired-end information, linked, and verified by mapping back the reads on the resulting assemblies. The overall coverage across each genome was then assessed to detect the presence of assembly artifacts potentially caused by repeated/duplicated regions differing from a 1:1 coverage ratio. To do so, reads were mapped on the final assembly with Bowtie 0.12.8 (32) and the assembly visually inspected with Tablet 1.12 (33). This strategy produced genomes with 1, 5, and 4 gaps in the 11 chromosome cores from ECI, ECII, and ECII, respectively, and these assemblies (with annotation [see below]), were used to update the initial draft releases.

Genome annotation. Genes coding for tRNAs were identified with tRNAscan-SE 1.21 (34), while rRNA-encoding genes were identified by BLAST homology searches using orthologs as input queries. The *E. cuniculi* GB-M1 protein-coding annotations were transferred on each of the *E. cuniculi* strain assemblies with RATT (35), with the start codons from the *E. cuniculi* GB-M1 EMBL annotation first reassessed as described by Pombert et al. (3). The curated *E. cuniculi* GB-M1 protein annotations were then transferred with RATT using the default parameters, and the annotations were verified with Artemis 14.0 (36). Genes missing from the transferred annotations were searched for specifically by BLAST homology searches using their *E. cuniculi* GB-M1 orthologs. Exon-intron junctions were verified manually.

Recombination analyses. Potential events of recombination between the three *E. cuniculi* strains were investigated on the colinear and conserved cores of each chromosome. Sequences from each chromosome core were aligned manually with BioEdit (version 7.1.3; Ibis Biosciences

[http://www.mbio.ncsu.edu/]), and recombination events were searched for with the RDP, GENECONV, BOOTSCAN, MaxChi, Chimera, SiScan, PhylPro, LARD, and 3Seq algorithms as implemented in RDP 4 beta 16 (37).

Global SNP calling. Global SNP assessments were performed using *E. cuniculi* GB-M1 as a reference. Read quality for each Illumina paired-end set was assessed with FastQC (version 0.10.1; Babraham Bioinformatics, Babraham Institute [http://www.bioinformatics.babraham.ac.uk]). Because most of the reads showed a significant drop in quality after the bp 60, all reads were filtered using a sliding-window quality approach with Sickle (Bioinformatics Core, University of California, Davis [https://github.com/najoshi/sickle]) under the default parameters. Read quality was then reassessed with FastQC for each filtered data set. The reads filtered with Sickle were concatenated as single forward and reverse FASTQ files for each *E. cuniculi* strain and mapped on the GB-M1 reference with SOAP2 2.20 (38) using the paired-end information with the minimum and maximum insert length flags (−m and −x) set to 0 and 600, respectively. The SOAP2 output was sorted using the bash shell command "sort −k8,8 −k9,9n output > sorted_output," and SNPs were called with SOAPsnp 1.03 (39) on the sorted output under the assumption of monoploidy and with the −z! option to specify the Sanger scoring scheme. The SOAPsnp output was then filtered using custom Perl scripts (all custom-made scripts are available from the authors upon request).

Genome alignment SNPs and sliding windows. SNPs were called on the aligned chromosome cores using custom Perl scripts. Briefly, each aligned sequence was put into its own array, with one nucleotide per element, and the corresponding elements queried for the presence of gaps, SNPs, or invariants. The results of these queries were put into their own .gaps, .snps, and .invar files and downstream analyses performed on these files. For sliding-window analyses, each aligned chromosome was queried again and the output written as single strings containing the binary characters 0 and 1 for the absence and presence of a SNP, respectively. Note that gaps were not considered SNPs in this analysis. The SNPs sliding windows were calculated from the binary strings.

Gene and codon SNPs. The *E. cuniculi* strain orthologous protein-encoding genes were aligned automatically with the L-INS-i algorithm from MAFFT (40). Gene SNPs were called using the same approach as described above for the genome alignment SNPs. From the produced .gaps, .snps, and .invar files, a total of 60 out of the 1,806 aligned genes contained one or more gaps. For each of these 60 genes, alignments were verified manually with Seaview 4 (41) and edited whenever required to preserve the codon frame. SNPs were called again on the verified alignments and downstream analyses performed on the corresponding files. For codon analyses, we used a similar approach in which each array element is a codon instead of a nucleotide. Again, queried elements were written sorted as gaps, SNPs, or invariants and the gap elements inspected again to detect potential scripting issues. Each codon containing one or more SNPs was queried against a hash representing the universal genetic code used by these species, and the synonymous and nonsynonymous changes were reported to the corresponding files. Functional categories were assigned using the KEGG web server (42).

Accession numbers. The sequence reads for the *E. cuniculi* ECI, ECII, and ECIII strains have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers SRX002289, SRX002285, and SRX002287, respectively. The initial assemblies were submitted to GenBank under accession numbers AEWD01000000, AEWQ01000000, and AEWR01000000 according to the Broad Institute policy on rapid data release. The ECI, ECII, and ECIII annotations used in this study are available in MicrosporidiaDB (43), and updated assemblies have been deposited in GenBank under accession numbers AEWD01, AEWQ01, and AEWR01.

RESULTS

***Encephalitozoon* strains share identical genome architecture and nearly identical gene content.** Complete genomes were sequenced and assembled from representatives of each of the three

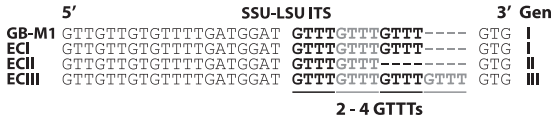

FIG 1 Genotypes of the four sequenced *E. cuniculi* strains as inferred from their small-subunit (SSU)–large-subunit (LSU) internal transcribed spacers. Genotypes (Gen), numbered as described by Didier et al. (14), are indicated on the right. The GB-M1 genotype was determined previously (1). Aside from their variable number of GTTTs, the four ITSs are identical.

recognized genotypes of *E. cuniculi*, ECI, ECII, and ECIII. These genotypes are defined by the number of GTTT repeats in the ITS region (14), so first the ITS was identified in each assembly, and the number of GTTT repeats was verified to correspond to the expectation for each genotype (three, two, and four GTTT copies, respectively, for ECI, ECII, and ECIII) (Fig. 1).

Comparing the chromosomal cores showed that the three genotypes share almost identical gene contents and gene arrangements. The ECI genome is nearly indistinguishable from the previously described *E. cuniculi* GB-M1 genome sequence, which also shares the same ITS repeats (Fig. 1). Only three minor differences between the ECI/ECII/ECIII coding contents were found (gene names are derived from locus tags from the GB-M1 annotation under GenBank accession numbers AL391737 and AL590442 to -51): (i) the gene ECU06_0740 is absent from ECII/ECIII genomes (it is also present in *E. intestinalis* but absent from *E. hellem* and *E. romaleae* and appears to be a distant paralog of ECU10_1480 found in all four congeners); (ii) in ECI the genes ECU06_0690 and ECU06_0700 are found in two distinct pieces, whereas in ECII and ECIII, as well as in *E. intestinalis*, *E. romaleae*, and *E. hellem*, these genes form a single open reading frame; and (iii) the ECIII genome contains only one of the three highly similar and adjacent paralogs ECU08_1700, ECU08_1710, and ECU08_1720 that are present in the ECI and ECII genomes (this same gene is found twice in *E. intestinalis* and once in *E. hellem* and *E. romaleae*). Altogether these are minor variations: the genomes share 1,857 other genes, all in a conserved order and orientation.

High levels of single-nucleotide polymorphisms in *Encephalitozoon cuniculi*. We used two different approaches to measure genome-wide levels of polymorphism among the *E. cuniculi* isolates. First, we detected SNPs by mapping the quality-filtered Illumina paired-end reads from each isolate against the GB-M1 genome sequence. Second, we identified SNPs from the aligned chromosome cores of the three *E. cuniculi* strains. Both approaches uncovered a large number of SNPs. The read-mapping approach yielded 757, 9,805, and 9,505 SNPs between GB-M1 and ECI, ECII, and ECIII, respectively. The chromosome alignments, which are more conservative, revealed 8,316, 8,061, and 2,208 SNPs between the ECI/ECII, ECI/ECIII, and ECII/ECIII pairs, 9,290 SNPs among all three strains, and 99 SNPs between ECI versus GB-M1 (Tables 1 and 2; Fig. 2). The difference in the number of SNPs identified by the different methods is substantial, especially those inferred between ECI and GB-M1. Nearly all of this difference can be attributed to the smaller subset used in the genome alignments: we restricted the chromosome alignments to the conserved chromosome cores and did not include the highly variable subtelomeric regions because the high levels of recent paralogy and possible intrastrain variation in subtelomeric regions would lead to exaggerated numbers of SNPs from compar-

TABLE 1 SNPs inferred from the aligned ECI, ECII, and ECIII chromosome cores

Chromosome[a]	Length (bp)		No. of SNPs				Gaps (bp)	Invariants (bp)	SNP density[c]
	Total	Gapless	ECI/ECII/ECIII	ECI/ECII	ECI/ECIII	ECII/ECIII			
I	135,208	133,992	608	536	492	188	0	133,384	4.5
II	176,356	174,825	722	649	617	178	87	174,103	4.1
III	180,860	179,206	784	703	677	188	86	178,422	4.3
IV	190,168	188,720	711	650	641	131	26	188,009	3.7
V	188,853	187,346	723	648	617	181	61	186,623	3.8
VIa	85,543	84,814	359	333	315	71	10	84,455	4.2
VIb	111,579	110,446	560	496	497	129	11	109,886	5.0
VII	207,703	205,652	899	816	783	199	253	204,753	4.3
VIII	210,797	207,213	960	867	858	195	1,664[b]	206,253	4.6
IXa	66,619	66,109	246	232	216	44	18	65,863	3.7
IXb	49,934	49,624	150	130	118	53	9	49,474	3.0
IXc	37,259	36,896	158	136	137	43	47	36,738	4.2
IXd	65,173	64,660	256	227	219	66	1	64,404	3.9
X	235,277	232,985	973	881	853	213	345	232,012	4.1
XI	250,043	247,569	1,181	1,012	1,021	329	112	246,388	4.7
Total	2,191,372	2,170,057	9,290	8,316	8,061	2,208	2,730	2,160,767	4.2

[a] See the Fig. 2 legend for a complete description of the aligned chromosome segments. Chromosomes for which portions could not be linked or aligned were broken into ordered segments (a, b, c, or d).
[b] The large number of gaps in chromosome VIII is caused by the absence of the first two of the three paralogs ECU08_1700, ECU08_1710, and ECU08_1720 in ECIII.
[c] Average number of SNPs per kb between all three strains (I.II.III).

ing nonorthologous genes. This and other pitfalls of SNP calling by read mapping (see reference 44 for details) lead us to favor the more conservative chromosome alignment approach, with the caveat that this does not consider possibly interesting variation in the subtelomeric regions. Given the similarity between the ECI and GB-M1 genomes by either analysis, we did not include GB-M1 in the tables or our downstream analyses (only two genes, ECU03_0290 and ECU03_1610, display SNPs between ECI and GB-M1).

The *E. cuniculi* strains have an average polymorphism density of 4.2 SNPs per kb. This density is consistent between chromosomes, ranging from 3.7 to 4.7 (Table 1; Fig. 2). ECII and ECIII isolates are more similar to each other than they are to ECI (Fig. 3). To assess the number of SNPs within coding regions, we aligned the 1,857 genes that are shared between the *E. cuniculi* chromosome cores: 46 tRNA-, 3 rRNA-, 2 U2 snRNA-, and 1,806 protein-coding genes. Most SNPs are in coding DNA (Table 2), which is not surprising given that they represent ~90% of the chromosomal regions that we investigated. The intergenic regions, which account for a small proportion of the *E. cuniculi* genome (~10%), harbor 5.6 SNPs per kb, which is slightly higher than the average for coding DNA (4.1 SNPs/kb).

A total of 255 genes were found to be invariant among the three isolates (see Table S1 in the supplemental material), including 67 involved in gene expression, 12 involved in purine/pyrimidine metabolism, RNA transport, and DNA repair, and 94 encoding hypothetical proteins. There were 1,604 genes with SNPs, which we ranked based on their level of divergence (see Table S2 in the supplemental material). Genes from the top 10th percentile code almost exclusively for hypothetical proteins, but we did find 8 genes coding for ribosomal proteins that had surprisingly high numbers of synonymous polymorphisms: ECU04_1355, ECU08_1910, ECU10_0160, ECU03_0710, ECU03_1490, ECU05_0920, ECU06_1445, and ECU08_1780. When only nonsynonymous changes are taken into account, however, these ribosomal genes are excluded from the top 10th percentile, suggesting that they are still under purifying selection.

A paucity of indels. Among the *E. cuniculi* isolates, there are 60 genes that contain insertions or deletions (indels), 16 of which alter the reading frame (see Table S3 in the supplemental material); however, in four of these 16 genes (ECU03_0680, ECU09_1850, ECU10_0690, and ECU11_0260) there are additional compensatory mutations that reestablish the reading frame. Only three of the 16 (ECU06_0100, ECU06_0700, and ECU09_1410) have indels that could be interpreted as likely

TABLE 2 SNPs located in coding regions between ECI, ECII, and ECIII

Genotypes	No. of SNPs							
	Total	Intergenic	Coding[a]	Protein[b]	Codon[c]	Synonymous	Nonsynonymous	Synonymous/nonsynonymous
ECI/ECII/ECIII	9,290	1,277	8,013	8,006	7,874	4,856	3,018	1.609
ECI/ECII	8,316	1,211	7,105	7,099	7,030	4,432	2,598	1.706
ECI/ECIII	8,061	1,093	6,968	6,964	6,848	4,311	2,537	1.699
ECII/ECIII	2,208	250	1,958	1,954	1,884	979	905	1.082

[a] Includes RNA- and protein-encoding genes.
[b] Protein-encoding genes only.
[c] Total number of distinct codons interrupted by SNPs.

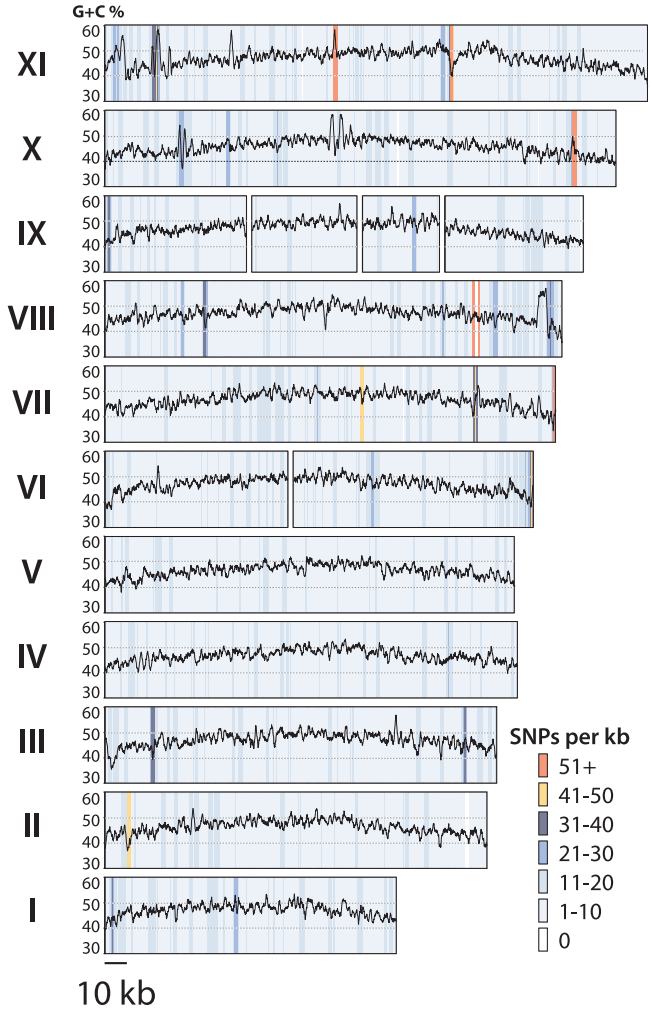

FIG 2 Occurrence of SNPs between ECI, ECII, and ECIII across their aligned chromosome cores. The SNP values across the three strains were calculated using a sliding window of 1,000 bp and a slide of 100 bp. The chromosomes are shown to scale from their 5′ to 3′ ends. The G+C content is plotted above using the same sliding-window parameters. The regions aligned for each chromosome are as follows: I (ECU01_0220 to ECU01_1390), II (ECU02_0090 to ECU02_1530), III (ECU03_0100 to ECU03_1610), IV (ECU04_0120 to ECU04_1625), V (ECU05_0060 to ECU05_1550), VIa (left; ECU06_0090 to ECU06_0730), VIb (right; ECU06_0750 to ECU06_1610), VII (ECU07_0080 to ECU07_1800), VIII (ECU08_0100 to ECU08_2060), IXa (left; ECU09_0020 to ECU09_0520), IXb (mid-left; ECU09_0600 to ECU09_1110), IXc (mid-right; ECU09_1170 to ECU09_1520), IXd (right; ECU09_1560 to ECU09_2000), X (ECU10_0150 to ECU10_1800), and XI (ECU11_0060 to ECU11_2037).

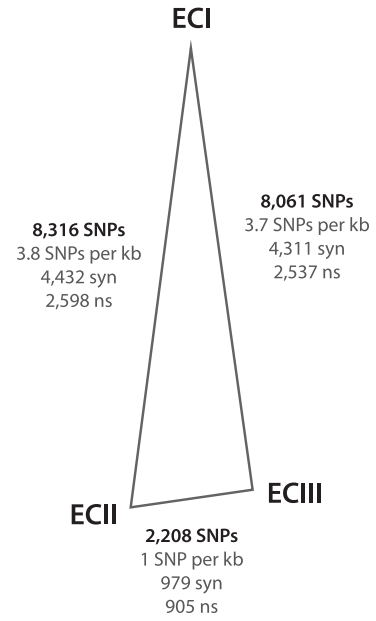

FIG 3 Genetic SNP distances between the ECI, ECII, and ECIII strains. Pairwise SNP distances between strains are shown adjacent to the corresponding triangle edges. The total number of SNPs between each pair is shown in bold. The total numbers of synonymous (syn) and nonsynonymous (ns) changes were inferred from the 1,806 aligned protein-encoding genes. Triangle edges are drawn to scale.

knocking out gene function, and interestingly, all of these are found in ECIII. The ECU08_1720 gene from ECIII is longer than its ECI/ECII homologs, which may explain why in ECIII this gene is carried as a single copy, whereas in ECI and II multiple paralogs are found.

No evidence for recombination between genotypes. Analyses of SNPs among ECI, ECII, and ECIII revealed no solid evidence for recombination. Applying several methods using RDP 4 revealed only three regions that could be indicative of recombination, but only one region had a strong E value (5.925E−41). This region encompasses ECU11_0880, a CTP synthase-encoding gene that is paralogous to ECU11_0480. While this might be taken to represent a recombination event between strains, it seems more likely to represent recombination between paralogs within one strain or even a cryptic assembly error. Otherwise, SNPs are distributed more or less evenly across the genomes, and ECII and ECIII are consistently more closely related to one another than either is to ECI, altogether suggesting that the ITS repeats do represent the genome and that the three identified genotypes are distinct populations.

Identification of potential high-resolution markers for strain identification. While the ITS does identify the three genotypes, other markers have greater variability and might allow greater resolution for strain identification. Among the most variable genes were those for spore wall protein 1 (SWP1) (ECU10_1660) and the polar tube protein 1 (PTP1) (ECU06_0250), but these are perhaps too variable to be used as genotyping tools. Indeed, both SWP1 and PTP1 differed between the two genotype 1 strains (GB-M1 and ECI) despite the fact that these strains had only 99 SNPs. In both cases, indels are found between the GB-M1, ECI, ECII, and ECIII strains (Fig. 4). Also, the ECII PTP1 sequence reported here differs from those previously reported (45). Because both proteins are antigens and therefore likely adapt rapidly to the host immune response, their observed diversity might not represent the genome as a whole particularly well. We searched for alternate molecular markers that are identical between ECI and GB-M1 but that can differentiate between ECI, ECII, and ECIII. We looked for genes that (i) display at least 7 SNPs per kb (see Table S2 in the supplemental material), (ii) are at least 1,000 bp long, (iii) are not paralogs, and (iv) have been attributed putative functions or show homology with conserved protein domains. We found 22 genes that fit these criteria and are therefore potentially useful for genotyping (Table 3). Of these, eukaryotic translation initiation factor 2 (ECU01_0700), translation elongation

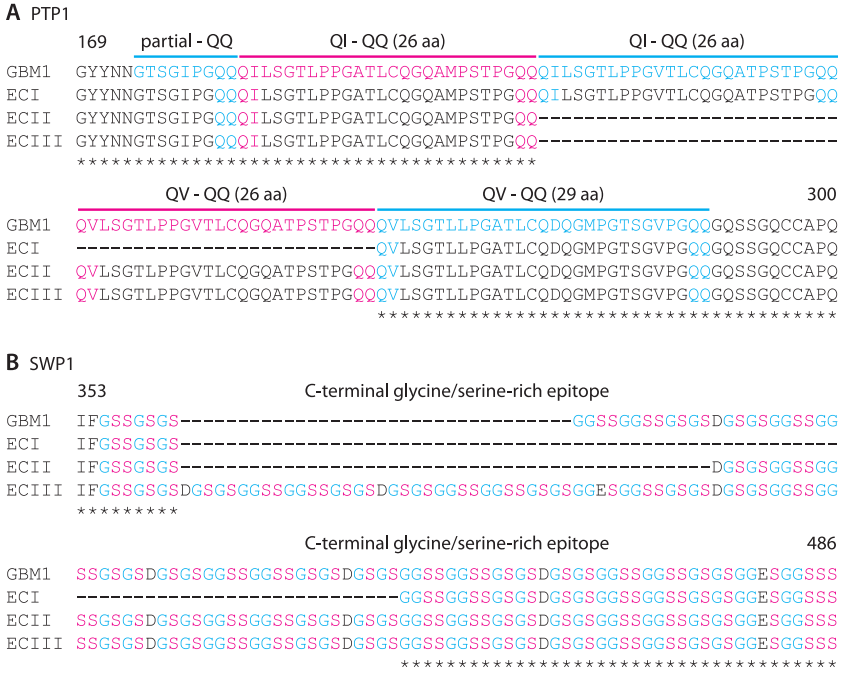

FIG 4 Amino acid (aa) alignments of the PTP1 (A) and SWP1 (B) variable regions between *E. cuniculi* strains. The repeated QI-QQ/QV-QQ motif in the polar tube protein 1 (PTP1) and the glycine/serine residues of the spore wall protein 1 (SWP1) C-terminal epitope are shown in alternating cyan/magenta colors. Gaps are denoted by dashes.

factor EF-1 alpha (ECU06_1440), and U2 snRNP/pre-mRNA association factor (ECU07_0340) may be particularly useful, as they are not involved in antigenic selection.

There are also eight *E. cuniculi* loci that have a particularly high SNP density that correlated with an upward or downward shift in G+C content compared to their immediate genomic surroundings (Fig. 5). Two of these loci include genes that are absent from other *Encephalitozoon* species (chromosome X, ECU10_0450 and ECU10_0460; chromosome XI, ECU11_0250 and ECU11_0260). It is not known if these genes were present in the ancestor of the *Encephalitozoon* genus and lost in certain species or if they were acquired in *E. cuniculi* through horizontal gene transfer. Horizontally transferred genes often have differing nucleotide contents and elevated mutation rates relative to their neighboring regions

TABLE 3 Genes of potential interest as genotyping tools in *E. cuniculi*

Gene	Length bp	No. of SNPs	No. of SNPs per kb	Product
ECU01_0450	1,191	11	9.2	DNA repair protein RAD4
ECU01_0700	1,320	10	7.6	Eukaryotic translation initiation factor 2 (eIF2)
ECU01_0830	1,041	9	8.6	CCCH-type Zn finger protein
ECU02_0330	2,922	21	7.2	Putative E1-E2 ATPase
ECU03_0560	1,011	9	8.9	Putative GTPase-activating protein
ECU03_0990	1,710	12	7.0	SCP/PR1 domain-containing protein
ECU04_1200	1,047	8	7.6	SWIB domain-containing protein
ECU04_1260	1,758	16	9.1	Nuclear protein export factor
ECU05_0220	2,346	17	7.2	WD40 domain-containing protein
ECU05_0240	1,533	11	7.2	Putative nitric oxide synthase
ECU06_1440	1,275	9	7.1	Translation elongation factor EF-1 alpha
ECU07_0340	1,002	8	8.0	U2 snRNP/pre-mRNA association factor
ECU07_0680	3,318	28	8.4	Chromosome segregation ATPase
ECU07_0830	3,177	33	10.4	Ski2-like helicase
ECU08_0330	1,083	8	7.4	Putative GTPase
ECU08_0400	1,548	11	7.1	tRNA/rRNA cytosine-C5-methylase
ECU08_1120	1,860	14	7.5	Rad3-like DNA helicase
ECU08_1770	1,017	9	8.8	DNA binding factor subunit TFIIIC1-like protein
ECU09_1850	1,155	9	7.8	PHD zinc finger domain-containing protein
ECU11_0350	1,092	8	7.3	Putative RAB escort protein
ECU11_0760	1,101	8	7.3	Putative exonuclease
ECU11_1540	1,371	11	8.0	Hexokinase

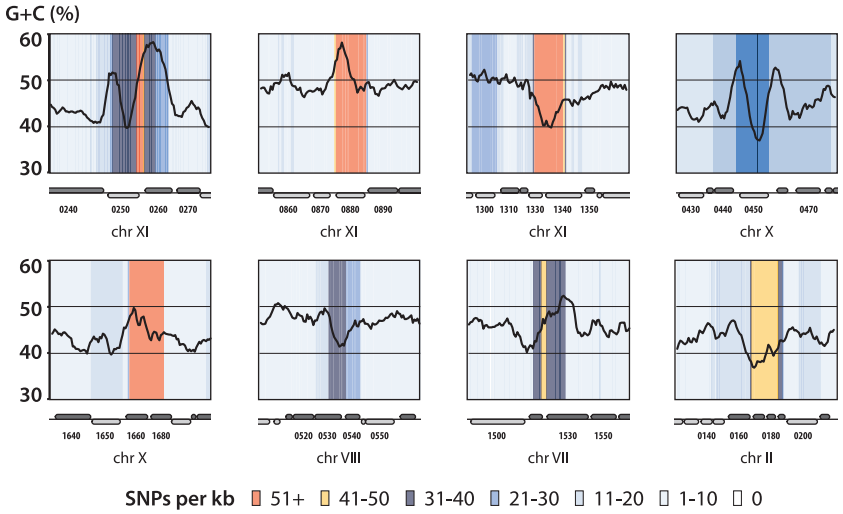

FIG 5 Cooccurrences of high SNP content and unusual G+C shifts in the *E. cuniculi* genomes. Each panel represent a 10-kb slice of the genome. The G+C plots are shown by black lines above the SNP content (color coded in the background). The genomic context is shown below each panel. Genes on the forward strand are shown in dark gray rounded rectangles. Genes on the lagging strand are shown in light gray rounded rectangles. The numbers above the chromosome labels indicate the corresponding *E. cuniculi* genes (e.g., chr XI + 0240 = ECU11_0240).

(see references 46 and 47 and references therein). The other six loci are found in all available *Encephalitozoon* genomes, where their nucleotide contents are also dissimilar to that of their surroundings in these other species, but the genes involved may simply belong to families with unusual G+C biases. Most of these genes code for proteins whose functions have yet to be determined, but interestingly, four (ECU11_0870, ECU11_1330, ECU10_0440, and ECU10_1680) are potentially involved in regulation of transcription. Unfortunately, it is not yet known whether there is a high SNP density at these loci in all *Encephalitozoon* species (because only a single strain each of *E. intestinalis*, *E. hellem*, and *E. romaleae* has been completely sequenced), which would allow us to distinguish whether the relationship between SNPs and G+C shifts in *E. cuniculi* is due to real biological processes or the effect of random noise.

Comparing diversity within *E. cuniculi* to diversity between *Encephalitozoon* species. We identified 203 protein-coding genes (36,445 codons) that are invariable among the three *E. cuniculi* isolates and used these genes to measure the diversity between the four *Encephalitozoon* species for which complete genomes are available. Of the 36,445 aligned codons, 28,310 (78%) had SNPs; 16,851 and 11,459 were synonymous and nonsynonymous, respectively. Overall, the invariable genes in *E. cuniculi* differ from their orthologs in *E. intestinalis*, *E. hellem*, and *E. romaleae* by 226, 222, and 221 SNPs per kb, giving an average nucleotide diversity of ~22%.

DISCUSSION

Knowing the levels of genetic diversity within and between populations is essential for understanding how organisms and genomes evolve, yet very little is known about the genetic diversity within microbial eukaryotes in general, and in the case of microsporidia, this gap is even more substantial. Indeed, it is not even known if microsporidia have sex (18, 48). In the case of *E. cuniculi*, various strains have been identified, and these are classified for diagnostic purposes as falling into three distinct genotypes based on the variable number of GTTT repeats in the ITS locus. However, whether the ITS locus adequately represents the entire genome had not been tested, and without genome-wide analyses, we cannot say for certain whether the three *E. cuniculi* genotypes come from the same or different populations/species.

Here, we analyzed complete sequences for representatives of each genotype and found no evidence for the exchange of genetic material among the three isolates, which means that the ITS is likely a reasonable representative of the genome as a whole. That said, the ITS regions did not capture the full extent of the diversity that we observed, and other markers, such as translation initiation factor 2 (ECU01_0700), translation elongation factor EF-1 alpha (ECU06_1440), and U2 snRNP/pre-mRNA association factor, might offer greater resolution between strains.

The genetic diversity values for the three *E. cuniculi* isolates are high compared to those of other microbial eukaryotic parasites but similar to those of other unicellular fungi. The average genetic distance among ECI, ECII, and ECIII (4.2 SNPs/kb) is 3 to 8 times those found among strains of *Plasmodium vivax* (0.8 SNPs/kb), *Plasmodium falciparum* (0.5 SNPs/kb), *Cryptosporidium parvum* (1.4 SNPs/kb), and *Entamoeba histolytica* (0.8 SNPs/kb) (49–51), but only 0.7 to 1.5 times those among free-living and pathogenic strains of *Saccharomyces cerevisiae* (2.8 and 6.1 SNPs/kb, respectively) (52, 53). Microsporidia are related to fungi; however, comparing diversity data from *E. cuniculi* to those from other fungi or other parasites is complicated by their unusual genomes: not only are the genes highly divergent, *Encephalitozoon* species also have the most compact nuclear genomes of all eukaryotes and consequently retain a small number of silent sites where SNPs might accumulate. Therefore, the vast majority of observed SNPs among the *E. cuniculi* strains occur in coding regions. Nevertheless, our analysis of the *E. cuniculi* invariable genes and their orthologs from the other *Encephalitozoon* species revealed a greater-than-50-fold divergence compared to the ECI/ECII/ECIII average genetic distance. This confirms the close relationship between the three *E. cuniculi* genotypes but also shows that the divergence is globally high in the Encephalitozoonidae. For example, the diver-

gence uncovered between the malaria parasites *P. falciparum* and *Plasmodium reichenowi* is about 1/10-fold lower, at 20 SNPs per kb (54).

The centromeres of microsporidian chromosomes have not yet been identified. It is known that the centromeres of yeast have a low G+C content (55, 56), display a high mutation rate (57), and are surrounded by regions of slightly lower polymorphism density than in the noncentromere regions (52). In *E. cuniculi*, however, all chromosomes display an arcing increase in G+C content from their edges to their center (Fig. 2), as previously observed by Katinka et al. (1), and there are only a few regions with unusual downward shifts in SNP content, most of which are located near the ends of the chromosomes (Fig. 2). The subtelomeric regions in microsporidia are known to evolve faster than the cores, and they also contain many paralogous genes that are likely involved in nonhomologous crossing-over events; this contrasts to the case for yeast chromosomes, which rarely undergo double-stranded breaks (58). Accordingly, whether the centromeres of *E. cuniculi* are located outside the chromosomal cores near the subtelomeric regions or whether they simply have different sequence characteristics than other genomes (which seems likely given the overall unusual nature of *Encephalitozoon* genomes) is uncertain.

The *E. cuniculi* genome does not appear to have a distinctive sex chromosome. In most eukaryotes with sex chromosomes, they have a lower SNP content than autosomes, but the three investigated genotypes share a similar SNP distribution among all chromosomes (Fig. 2). Microsporidia have been argued to possess a locus similar to the zygomycetes *sex* locus (48), a small genomic region restrained to only a few genes. Previously the levels of polymorphism at the putative *sex*-related genes were analyzed and shown to be minimal (59). Here, we find the polymorphism over the entire putative *sex*-related locus is on average roughly half that of the genome as a whole (2.4 versus 4.2 SNPs per kb, respectively), and the two high-mobility group (HMG) proteins (ECU06_1260 and ECU06_1270) located within it do not contain any SNPs, consistent with the previous finding that the only SNPs in this locus were positioned toward the edges of the region (59).

Conclusions. Comparative genomics has told us much about the extreme nature of microsporidian genomes, particularly those of the genus *Encephalitozoon*. They have not, however, shed much light on the actual processes that led to their extreme states, because comparisons have been restricted to distant relations, at best between two congeneric species (18). To begin to observe the processes at work within these genomes, a new level of comparison within populations will be most informative, as long as we can accurately identify genetically distinct strains and there is sufficient genetic diversity within them. Here we show that both of these are realistic expectations for a model species, *E. cuniculi*. As described in an accompanying paper, comparisons between genomes of three ITS-defined genotypes confirm they are genetically distinct and indeed that they have a high level of polymorphism despite extremely low levels of heterozygosity between alleles (60). While this already reveals a number of interesting trends in the population structure of this lineage, it also suggests that future studies using large-scale population genomics approach will be extremely helpful in explaining how microsporidian genomes actually evolve at a fine scale.

ACKNOWLEDGMENTS

This work was supported by grants from the Canadian Institute of Health Research to P.J.K. (MOP-42517) and from the National Institutes of Health to L.M.W. (NIAID AI31788). This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN272200900018C. P.J.K. is a Fellow of the Canadian Institute for Advanced Research (CIFAR).

We gratefully acknowledge the Compute Canada/CLUMEQ consortium for access to their computational resources.

REFERENCES

1. **Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP.** 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature **414**:450–453.
2. **Tsaousis AD, Kunji ERS, Goldberg AV, Lucocq JM, Hirt RP, Embley TM.** 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. Nature **453**:553–556.
3. **Pombert J-F, Selman M, Burki F, Bardell FT, Farinelli L, Solter LF, Whitman DW, Weiss LM, Corradi N, Keeling PJ.** 2012. Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. Proc. Natl. Acad. Sci. U. S. A. **109**:12638–12643.
4. **Corradi N, Pombert Farinelli J-FL, Didier ES, Keeling PJ.** 2010. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. Nat. Commun. **1**:77.
5. **Santín M, Fayer R.** 2009. *Enterocytozoon bieneusi* genotype nomenclature based on the internal transcribed spacer sequence: a consensus. J. Eukaryot. Microbiol. **56**:34–38.
6. **Li W, Cama V, Feng Y, Gilman RH, Bern C, Zhang X, Xiao L.** 2012. Population genetic analysis of *Enterocytozoon bieneusi* in humans. Int. J. Parasitol. **42**:287–293.
7. **Widmer G, Akiyoshi DE.** 2010. Host-specific segregation of ribosomal nucleotide sequence diversity in the microsporidian *Enterocytozoon bieneusi*. Infect. Genet. Evol. **10**:122–128.
8. **Santín M, Fayer R.** 2011. Microsporidiosis: *Enterocytozoon bieneusi* in domesticated and wild animals. Res. Vet. Sci. **90**:363–371.
9. **Li J, Chen W, Wu J, Peng W, An J, Schmid-Hempel P, Schmid-Hempel R.** 2012. Diversity of *Nosema* associated with bumblebees (*Bombus* spp.) from China. Int. J. Parasitol. **42**:49–61.
10. **Sagastume S, Del Aguila C, Martín-Hernández R, Higes M, Henriques-Gil N.** 2011. Polymorphism and recombination for rDNA in the putatively asexual microsporidian *Nosema ceranae*, a pathogen of honeybees. Environ. Microbiol. **13**:84–95.
11. **Surendra Nath B, Hassan W, Nageswara Rao S, Vijaya Prakash NB, Gupta SK, Madana Mohanan N, Bajpai AK.** 2011. Genetic diversity among microsporidian isolates from the silkworm, *Bombyx mori*, as revealed by randomly amplified polymorphic DNA (RAPD) markers. Acta Parasitol. **56**:333–338.
12. **Haro M, Del Aguila C, Fenoy S, Henriques-Gil N.** 2003. Intraspecies genotype variability of the microsporidian parasite *Encephalitozoon hellem*. J. Clin. Microbiol. **41**:4166–4171.
13. **Lange CE, Johny S, Baker MD, Whitman DW, Solter LF.** 2009. A new Encephalitozoon species (Microsporidia) isolated from the lubber grasshopper, *Romalea microptera* (Beauvois) (Orthoptera: Romaleidae). J. Parasitol. **95**:976–986.
14. **Didier ES, Vossbrink CR, Baker MD, Rogers LB, Bertucci DC, Shadduck JA.** 1995. Identification and characterization of three *Encephalitozoon cuniculi* strains. Parasitology **111**:411–421.
15. **Mathis A, Tanner I, Weber R, Deplazes P.** 1999. Genetic and phenotypic intraspecific variation in the microsporidian *Encephalitozoon hellem*. Int. J. Parasitol. **29**:767–770.
16. **Polonais V, Mazet M, Wawrzyniak I, Texier C, Blot N, El Alaoui H, Delbac F.** 2010. The human microsporidian *Encephalitozoon hellem* synthesizes two spore wall polymorphic proteins useful for epidemiological studies. Infect. Immun. **78**:2221–2230.
17. **Xiao L, Li L, Moura H, Sulaiman I, Lal AA, Gatti S, Scaglia M, Didier**

ES, Visvesvara GS. 2001. Genotyping *Encephalitozoon hellem* isolates by analysis of the polar tube protein gene. J. Clin. Microbiol. **39**:2191–2196.
18. **Cuomo CA, Desjardins CA, Bakowski MA, Goldberg J, Ma AT, Becnel JJ, Didier ES, Fan L, Heiman DI, Levin JZ, Young S, Zeng Q, Troemel ER.** 2012. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. **22**:2478–2488.
19. **Didier ES, Didier PJ, Snowden KF, Shadduck JA.** 2000. Microsporidiosis in mammals. Microbes Infect. **2**:709–720.
20. **Shadduck JA.** 1969. *Nosema cuniculi*: *in vitro* isolation. Science **166**:516–517.
21. **Vávra J, Bedrník P, Cinatl J.** 1972. Isolation and *in vitro* cultivation of the mammalian microsporidia *Encephalitozoon cuniculi*. Folia Parasitol. **19**:349–354.
22. **Shadduck JA, Bendele R, Robinson GT.** 1978. Isolation of the causative organism of canine encephalitozoonosis. Vet. Pathol. **15**:449–460.
23. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
24. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760.
25. **Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.** 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. **108**:1513–1518.
26. **Sommer DD, Delcher AL, Salzberg SL, Pop M.** 2007. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics **8**:64.
27. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.
28. **Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I.** 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. **19**:1117–1123.
29. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5**:R12.
30. **Boisvert S, Laviolette F, Corbeil J.** 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. **17**:1519–1533.
31. **Gordon D, Abajian C, Green P.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8**:195–202.
32. **Langmead B, Trapnell C, Pop M, Salzberg SL.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**:R25.
33. **Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D.** 2010. Tablet—next generation sequence assembly visualization. Bioinformatics **26**:401–402.
34. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.
35. **Otto TD, Dillon GP, Degrave WS, Berriman M.** 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. **39**:e57.
36. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16**:944–945.
37. **Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P.** 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26**:2462–2463.
38. **Li R, Yu C, Li Y, Lam T-WYiu Kristiansen S-MK, Wang J.** 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25**:1966–1967.
39. **Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J.** 2009. SNP detection for massively parallel whole-genome resequencing. Genome Res. **19**:1124–1132.
40. **Katoh K, Toh H.** 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics **26**:1899–1900.
41. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. **27**:221–224.
42. **Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M.** 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. **40**:D109–D114.
43. **Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Iodice J, Kissinger JC, Kraemer ET, Li W, Nayak V, Pennington C, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ, Treatman C, Wang H.** 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. Nucleic Acids Res. **39**:D612–D619.
44. **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**:491–498.
45. **Peuvel I, Delbac F, Metenier G, Peyret P, Vivares CP.** 2000. Polymorphism of the gene encoding a major polar tube protein PTP1 in two microsporidia of the genus *Encephalitozoon*. Parasitology **121**:581–587.
46. **Syvanen M.** 2012. Evolutionary implications of horizontal gene transfer. Annu. Rev. Genet. **46**:339–356.
47. **Keeling PJ, Palmer JD.** 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. **9**:605–618.
48. **Lee SC, Weiss LM, Heitman J.** 2009. Generation of genetic diversity in microsporidia via sexual reproduction and horizontal gene transfer. Commun. Integr. Biol. **2**:414–417.
49. **Neafsey DE, Galinsky K, Jiang RHY, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM.** 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat. Genet. **44**:1046–1050.
50. **Weedall GD, Clark CG, Koldkjaer P, Kay S, Bruchhaus I, Tannich E, Paterson S, Hall N.** 2012. Genomic diversity of the human intestinal parasite *Entamoeba histolytica*. Genome Biol. **13**:R38.
51. **Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K.** 2012. Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range. Infect. Genet. Evol. **12**:1213–1221.
52. **Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L.** 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature **458**:342–345.
53. **Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y, Davis RW, Steinmetz LM.** 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proc. Natl. Acad. Sci. U. S. A. **104**:12825–12830.
54. **Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, Ingle CE, Thomas A, Quail MA, Siebenthall K, Uhlemann Kyes A-CS, Krishna S, Newbold C, Dermitzakis ET, Berriman M.** 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. Nat. Genet. **39**:120–125.
55. **Wendland J, Walther A.** 2011. Genome evolution in the *Eremothecium* clade of the *Saccharomyces complex* revealed by comparative genomics. G3 (Bethesda) **1**:539–548.
56. **Lynch DB, Logue ME, Butler G, Wolfe KH.** 2010. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol. Evol. **2**:572–583.
57. **Bensasson D.** 2011. Evidence for a high mutation rate at rapidly evolving yeast centromeres. BMC Evol. Biol. **11**:211.
58. **Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD.** 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U. S. A. **97**:11383–11390.
59. **Lee SC, Corradi N, Doan S, Dietrich FS, Keeling PJ, Heitman J.** 2010. Evolution of the *sex*-related locus and genomic features shared in microsporidia and fungi. PLoS One **5**:e10539.
60. **Selman M, Sak B, Kvác M, Farinelli L, Weiss LM, Corradi N.** 2013. Extremely reduced levels of heterozygosity in the vertebrate pathogen *Encephalitozoon cuniculi*. Eukaryot. Cell **12**:522–528.