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

Jean-François Pombert,*¹ Jinshan Xu,*¹ David R. Smith,*¹ David Heiman,*¹ Sarah Young,*¹ Christina A. Cuomo,*¹ Louis M. Weiss,*¹ Patrick J. Keeling²

Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada*¹, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA², Department of Pathology, Division of Parasitology, and Department of Medicine, Division of Infectious Diseases, Albert Einstein College of Medicine, Bronx, New York, USA*¹

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
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 these 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% CO2 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 H2O, 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 debris, and the supernatants were transferred with RATT using the default parameters, and the raw data were then filtered against the source organism. A contaminating sequence or identify reads matching the source organism. A 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.ua.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 ECIII, 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 mRNA-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 (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)
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 –k9n 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, .snp, 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 SNP’s sliding windows were calculated from the binary strings.

Gene and codon SNPs. The E. cuniculi strain orthologous protein-coding 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, .snp, 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 SNP’s sliding windows were calculated from the binary strings.

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/ECI/ECII 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 ECI/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 ECII genome contains only one of the three highly similar and adjacent paralogs 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 intraspecies variation in subtelomeric regions would lead to exaggerated numbers of SNPs from compar-
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 ECII 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, ECU06_0100, ECU06_0700, and ECU09_1410) have indels that could be interpreted as likely

| TABLE 1 SNPs inferred from the aligned ECI, ECII, and ECIII chromosome cores |
|-----------------------------|-------------------|-------------------|-------------------|-------------------|
| Chromosome               | Length (bp) | Gapless total | ECI/ECII/ECIII | ECI/ECII | ECI/ECIII | ECI/ECIII | Gaps (bp) | Invariants (bp) | SNP density |
| I                          | 135,208     | 133,992        | 608            | 536     | 492       | 188       | 133,384    | 4.5          |
| II                         | 176,356     | 174,825        | 722            | 649     | 617       | 178       | 174,103    | 4.1          |
| III                        | 180,860     | 179,206        | 784            | 703     | 677       | 188       | 178,422    | 4.3          |
| IV                         | 190,168     | 188,720        | 711            | 650     | 641       | 131       | 186,009    | 3.7          |
| V                          | 188,853     | 187,346        | 723            | 648     | 617       | 181       | 186,623    | 3.8          |
| VIa                       | 85,543      | 84,814         | 359            | 333     | 315       | 71        | 84,455     | 4.2          |
| VIb                       | 111,579     | 110,446        | 560            | 496     | 497       | 129       | 109,886    | 5.0          |
| VII                        | 207,703     | 205,652        | 899            | 816     | 783       | 199       | 204,753    | 4.3          |
| VIII                      | 210,797     | 207,213        | 960            | 867     | 858       | 195       | 206,253    | 4.6          |
| IXa                       | 66,619      | 66,109         | 246            | 232     | 216       | 44        | 65,863     | 3.7          |
| IXb                       | 49,934      | 49,624         | 150            | 130     | 118       | 53        | 49,474     | 3.0          |
| IXc                       | 37,259      | 36,896         | 158            | 136     | 137       | 43        | 36,738     | 4.2          |
| IXd                       | 235,277     | 232,985        | 973            | 881     | 853       | 213       | 232,012    | 4.1          |
| XI                        | 250,043     | 247,569        | 1,181          | 1,012   | 1,021     | 329       | 246,388    | 4.7          |
| Total                     | 2,191,372   | 2,170,057      | 9,290          | 8,316   | 8,061     | 2,208     | 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.

Average number of SNPs per kb between all three strains (I.II.III).

| TABLE 2 SNPs located in coding regions between ECI, ECII, and ECIII |
|-----------------------------|-------------------|
| Genotypes                  | No. of SNPs |
| ECI/ECII/ECIII             | 9,290          |
| ECI/ECII                   | 8,316          |
| ECI/ECIII                  | 8,061          |
| ECI/ECII/ECIII             | 2,208          |
| Coding genotypes           | 8,013          |
| Protein genotypes          | 8,006          |
| Codon genotypes            | 7,874          |
| Synonymous                 | 4,856          |
| Nonsynonymous              | 3,018          |
| Syonymous/nonsynonymous    | 1,609          |

a Includes RNA- and protein-encoding genes.

b Protein-encoding genes only.

c Total number of distinct codons interrupted by SNPs.

Pombert et al.
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/H1100241). 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...
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 \textit{E. cuniculi} loci that have a particularly high SNP density that correlated with an upward or downward shift in $G$/$H_11001$ content compared to their immediate genomic surroundings (Fig. 5). Two of these loci include genes that are absent from other \textit{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 \textit{Encephalitozoon} genus and lost in certain species or if they were acquired in \textit{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 \textit{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               | Skil2-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   | 1,869     | 14          | 7.5                | Rad3-like DNA helicase                         |
| ECU08_1770   | 1,017     | 9           | 8.8                | DNA binding factor subunit TFIIC1-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                                    |
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* variable 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.
gence uncovered between the malaria parasites \textit{P. falciparum} and \textit{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 \textit{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). According, whether the centromeres of \textit{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 \textit{Encephalitozoon} genomes) is uncertain.

The \textit{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 zygomyces 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 ends of the region (59).

Conclusions. Comparative genomics has told us much about the extreme nature of microsporidian genomes, particularly those of the genus \textit{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 congenic 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, \textit{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. (NIH A131788). 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, Peyretaille E, Brotier P, Wincker P, Delbac F, El Aloui H, Peyret P, Saurin W, Gouy M, Weisschenbach J, Vivares CP. 2001. Genome sequence and gene compaction of the eukaryote parasite \textit{Encephalitozoon cuniculi}. Nature 414:450–453.
2. Tsaoissis AD, Kunji ERS, Goldberg AV, Lucoq JM, Hirt RP, Embley TM, 2008. A novel route for ATP acquisition by the remnant mitochondrial of \textit{Encephalitozoon cuniculi}. Nature 453:553–556.
3. Pombert J-F, Selman J, Burki F, Bardell FT, Farinelli L, Solfer L, 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 \textit{Encephalitozoon intestinalis}. Nat. Commun. 1:77.
5. Santín M, Fayer R. 2009. \textit{Enterocytozoon bieneusi} beneu gene 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 \textit{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 \textit{Enterocytozoon bieneusi}. Infect. Genet. Evol. 10:122–128.
8. Santín M, Fayer R. 2011. Microsporidiosis: \textit{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 \textit{Noema} 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 \textit{Noema ceraeae}, a pathogen of honeybees. Environ. Microbiol. 13:84–95.
11. Sundreth Nath B, Hassan W, Nageswara Rao S, Vijaya Prakash NB, Gupta SK, Madana Mohanan N, Baijai AK. 2011. Genetic diversity among microsporidian isolates from the silkworm, \textit{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 \textit{Encephalitozoon hellem}. J. Clin. Microbiol. 41:4166–4171.
13. Lange CE, Johny S, Baker MD, Whitman DW, Solfer LF. 2009. A new \textit{Encephalitozoon} species (Microsporida) isolated from the lubber grasshopper, \textit{Romaulea microptera} (Beauvios) (Orthoptera: Romaleidae). J. Parasitol. 95:976–986.
14. Didier ES, Vossbrink CR, Baker MD, Rogers LB, Bertucci DG, Shadwick JA, 1995. Identification and characterization of three \textit{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 \textit{Encephalitozoon hellem}. Int. J. Parasitol. 29:767–770.
16. Polonais V, Mazet M, Wawrzyniak I, Texier C, Buet N, El Aloui H, Delbac F. 2010. The human microsporidian \textit{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
