nature COMMUNICATIONS

ARTICLE

Received 1 Jun 2010 | Accepted 25 Aug 2010 | Published 21 Sep 2010

DOI: 10.1038/ncomms1082

The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*

Nicolas Corradi[1,*,†], Jean-François Pombert[1,*], Laurent Farinelli[2], Elizabeth S. Didier[3] & Patrick J. Keeling[1]

The genome of the microsporidia *Encephalitozoon cuniculi* is widely recognized as a model for extreme reduction and compaction. At only 2.9 Mbp, the genome encodes approximately 2,000 densely packed genes and little else. However, the nuclear genome of its sister, *Encephalitozoon intestinalis*, is even more reduced; at 2.3 Mbp, it represents a 20% reduction from an already severely compacted genome, raising the question, what else can be lost? In this paper, we describe the complete sequence of the *E. intestinalis* genome and its comparison with that of *E. cuniculi*. The two species share a conserved gene content, order and density over most of their genomes. The exceptions are the subtelomeric regions, where *E. intestinalis* chromosomes are missing large gene blocks of sequence found in *E. cuniculi*. In the remaining gene-dense chromosome 'cores', the diminutive intergenic sequences and introns are actually more highly conserved than the genes themselves, suggesting that they have reached the limits of reduction for a fully functional genome.

[1] Department of Botany, Canadian Institute for Advanced Research, University of British Columbia, 3529-6270 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z4. [2] FASTERIS S.A., Ch. du Pont-du-Centenaire 109, PO Box 28, CH-1228 Plan-les-Ouates, Switzerland. [3] Tulane National Primate Research Center, Tulane University, 18703 Three Rivers Road, Covington, Louisiana 70433, USA. *These authors contributed equally to this work. †Present address: Department of Biology, Canadian Institute for Advanced Research, University of Ottawa, Gendron Hall, Ottawa, Ontario, Canada K1N 6N5. Correspondence and requests for materials should be addressed to P.J.K. (email: pkeeling@interchange.ubc.ca).

Microsporidia are a group of obligate intracellular parasites of agricultural and medical importance that are widely recognized as highly adapted fungi[1–6]. Their obligate intracellular lifestyle is characterized by a high degree of host dependence[7], leading to, among other things, an extraordinary reduction in the number of genes encoded in their genomes. In addition to losing many genes, however, several microsporidian genomes have also evolved a very high gene density, partly by the shortening of the genes themselves, but more substantially by reducing their intergenic regions. In the most extreme cases, genes are tightly packed with intergenic spaces averaging just over 100 bp, and several protein-coding sequences physically overlap with their neighbours. The miniaturization of these genomes has affected not just form but also function, in particular leading to frequent overlaps between the mRNA transcripts of adjacent genes in many species[8–11]. Genome compaction has also seemingly affected the rate at which microsporidian parasites shuffle their genomes; hence, the order of genes is conserved, even between species separated by very large genetic distances[12,13], despite the fact that the genes themselves are known to be evolving very rapidly at the sequence level. Overall, microsporidian nuclear genomes are the most reduced and compacted of any eukaryotic cell, including picoplankton and other obligately intracellular parasites such as *Plasmodium*, the agent of malaria.

The model organism for these highly compacted genomes is the human parasite *Encephalitozoon cuniculi*, the completely sequenced genome of which is only 2.9 Mbp[3]. With approximately 2,000 genes,

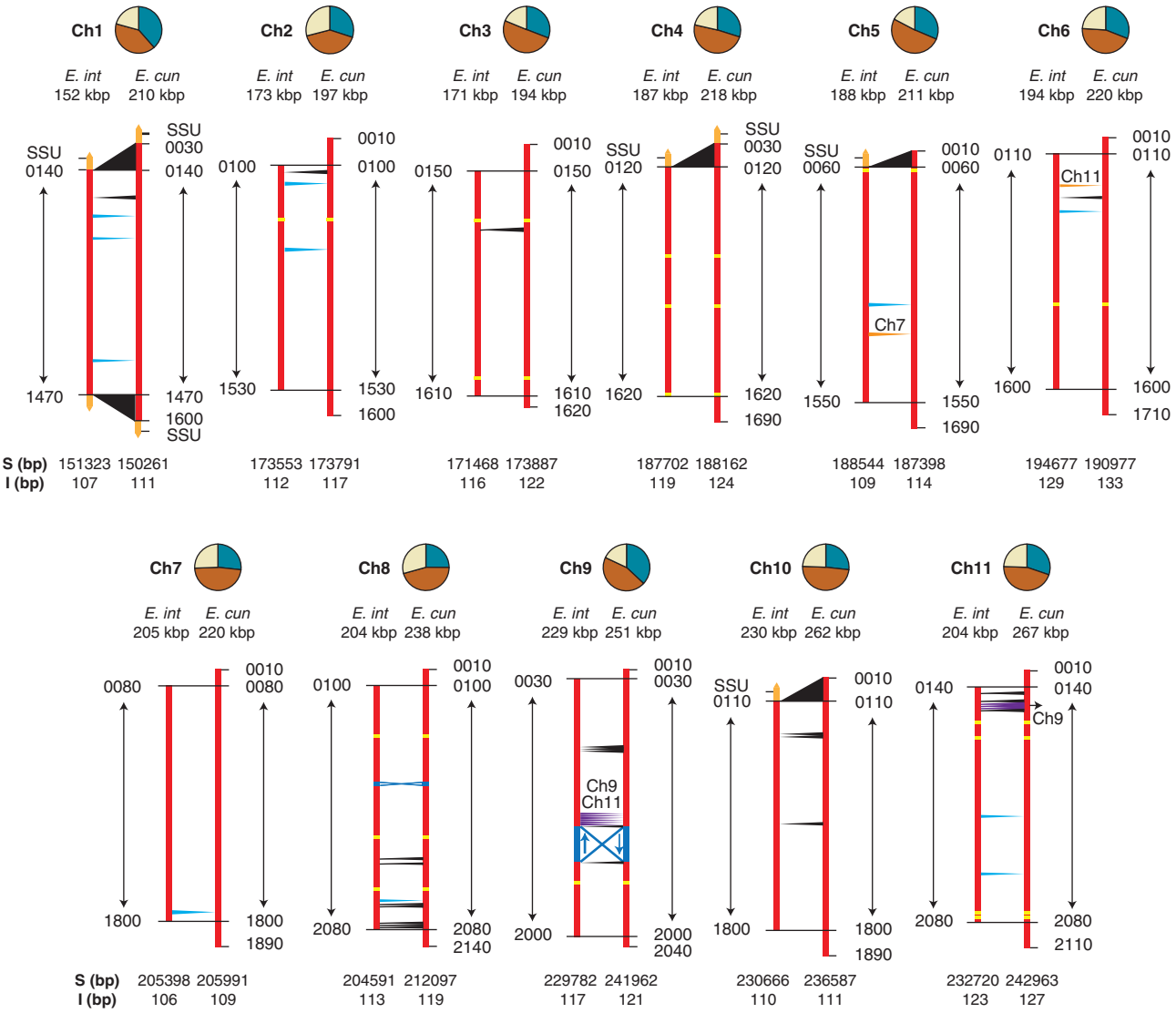

Figure 1 | Comparison between the chromosomes of *E. intestinalis* and *E. cuniculi*. Comparison of the 11 chromosomes of *E. intestinalis* (left side) and *E. cuniculi* (right side), with the total assembled size for each indicated below the name. Difference in the relative length of orthologous protein-coding genes is summarized in a pie chart above each chromosome. Blue, brown and beige colours represent the portion of proteins that are, respectively, shorter, identical or longer in *E. intestinalis* compared with *E. cuniculi* orthologues. Chromosome 'cores' are shown in red and the size (S) and average intergenic regions (I) of each core are indicated under it. Gene rearrangements, inversions and events of gene losses and gains between species are shown as coloured triangles. Black triangles represent the location of genes absent from *E. intestinalis*. Light blue rectangles represent the location of genes absent from *E. cuniculi*. Yellow rectangles represent genes that were previously unannotated in *E. cuniculi* that have been identified by comparisons with *E. intestinalis*. In addition, it was evident for many other genes that the previous annotation used the wrong ATG codon. The newly annotated version of the *E. cuniculi* genome is available as Supplementary Data 1. Dark orange triangles represent genes duplicated and rearranged between chromosomes of *E. intestinalis* (chromosome number shown above the rectangle). Dark violet arrows represent genes transposed from another chromosome (original chromosome number shown above the rectangle). Chromosomal inversions are shown in dark blue. SSU, small subunit ribosomal RNA gene.

Table 1 | General characteristics of the genomes of *E. intestinalis* and other microsporidia.

	Encephalitozoon intestinalis	*Encephalitozoon cuniculi**	*Enterocytozoon bieneusi*	*Octosporea bayeri*
Chromosomes (#)	11	11	≥6	Unknown
Genome size (Mbp)	2.3	2.9	6	≤24.2
Assembled (Mbp)	2.2	2.5	3.86	13.3
Genome coverage (%)	96	86	64	55
G+C content (%)	41.4	47	25	26
Gene density (gene per kbp)	0.86	0.84	0.87	0.23
Mean intergenic length	115 bp	119 bp	127 bp	429 bp
Presence of overlapping genes	Yes	Yes	Yes	No
SSU-LSU rRNA genes	22	22	Unknown	≥2
5S rRNA genes	3	3	Unknown	≥2
tRNAs	46	46	46	37
tRNA synthetases	21	21	21	21
tRNA introns (size)	2 (16, 42 bp)	2 (16, 42 bp)	2 (13, 30 bp)	≥1 (50 bp)
Spliceosomal introns (size)	14 (23–47 bp)	14 (23–49 bp)	0	≥6 (24–33 bp)
Predicted ORFs	1,833	1,999	3,804	2,174
ORFs with assigned functional categories	886 (48%)	894 (45%)	669 (39%)	894 (41%)
Mean size of CDS	1,041 bp	1,041 bp	1,002 bp	1,056 bp

Abbreviations: CDS, coding sequences; LSU, large subunit ribosomal RNA gene; ORF, open reading frames; rRNA: ribosomal RNA; SSU, small subunit ribosomal RNA gene; tRNA, transfer RNA.
*Values for *E. cuniculi* differ from those reported in the genome because our reannotation of the genome based on *E. intestinalis* (see Supplementary Data 1) altered the previous annotation of several *E. cuniculi* genes. Values for *E. bieneusi* and *O. bayeri* have been previously reported elsewhere[16,18].

this genome is indeed strikingly reduced; however, this is not the smallest known microsporidian genome. The genome of the closely related species, *E. intestinalis,* has been estimated by pulsed field gel electrophoresis to be only 2.3 Mbp[14], which corresponds to 600 kbp or 20% smaller than *E. cuniculi*. Given the already reduced nature of *E. cuniculi*, one wonders what else can be lost. At the extremes, either hundreds of genes were lost or the entire genome was even more radically compacted. Unfortunately, currently, no data exist on the nature of this smallest known nuclear genome to reveal what evolutionary forces might operate at the far end of the spectrum of genome reduction. In this study, we describe the complete genome sequence of *E. intestinalis* (ATCC 50506). A comparison between this genome and that of its sibling species, *E. cuniculi*, reveals that virtually all of the difference in genome size can be attributed to large subtelomeric regions that are present in *E. cuniculi* but absent from *E. intestinalis*. The remainder of the genome is relatively conserved in content, order and density, and indeed we find that the intergenic regions and the introns are remarkably well conserved at the sequence level, altogether suggesting that these chromosome 'cores' have reached a certain limit of reduction and additional substantial changes to them are likely to be difficult.

Results

General features of the *E. intestinalis* genome. A total of 200 ng of *E. intestinalis* DNA was isolated from purified spores and used for Solexa sequencing, from which the entire genome was assembled *de novo*, resulting in an assembly of 137 scaffolds with an average coverage of 40×. PCR and Sanger sequencing were used to link the preliminary scaffolds and polish internal breaks, resulting in 11 large contigs homologous to the 11 chromosomes of *E. cuniculi,* with a total sequence of 2,191,783 Mbp, or >95% of the estimated 2.3 Mbp (Fig. 1). The *E. intestinalis* genome harbours the same complement of transfer RNAs and ribosomal RNAs as *E. cuniculi* and both genomes have a similar GC content (Table 1). The subtelomeric regions of microsporidian chromosomes typically consist of a mixture of unique genes and repeated gene families, ending with a copy of the ribosomal RNA operon[3,15]. We found a 22-fold excess of ribosomal RNA operon in our assembly, suggesting that it is also located in the subtelomeric regions of all 11 chromosomes in *E. intestinalis*. In five cases, we linked the ribosomal RNA operon to the end of a chromosome assembly by PCR, representing all three subtelomeric regions known in *E. cuniculi*, plus two additional ones (Fig. 1).

Coding capacity of the smallest nuclear genome. The coding capacity and structure of the *E. intestinalis* genome were identified using manual annotation and compared against our own manually annotated version of the *E. cuniculi* genome (available as Supplementary Data 1) and the genomes of other microsporidian relatives[3,13,16–19]. Altogether, 1,833 protein-coding genes were identified in the assembly. If complete, this would result in a coding capacity that is almost 10% smaller than that of *E. cuniculi*, and the smallest identified in a eukaryote. The complexity of its proteome is, however, predicted to be very similar to that of other microsporidia (Table 1, Supplementary Table S1) because the majority of those genes found to be absent from *E. intestinalis* are hypothetical proteins or duplicates of genes that are present. Only 15 protein-coding genes with known function in *E. cuniculi* were found to be absent in *E. intestinalis,* whereas only four such genes were found in *E. intestinalis* but not in *E. cuniculi*. These did not correspond to whole pathways, but were instead a scattered representation of various functions (Supplementary Data 2). Interestingly, 10 other hypothetical genes with no known function were also found in *E. intestinalis* but have no recognizable orthologue in *E. cuniculi*; three of these are known from other microsporidia, whereas seven are exclusive to *E. intestinalis* (Supplementary Data 2). Finally, 16 genes that were previously unannotated in *E. cuniculi* were identified on the basis of homology to *E. intestinalis* (Supplementary Data 2), bringing the total number of protein-coding genes in *E. cuniculi* to 1,999 (Table 1). This number, which is based on a direct comparison with that of *E. intestinalis*, is lower than previous estimates, even though several new genes were identified because a number of previously annotated open reading frames (ORFs) now seem to be spurious (see our own annotation of the *E. cuniculi* genome, available as Supplementary Data 1).

Genome evolution along the chromosomes of *E. intestinalis*. A comparison between *E. intestinalis* and *E. cuniculi* genomes revealed that their chromosome 'cores' are, with a few exceptions, completely colinear, despite significant divergence in sequence (Fig. 1; see 'Methods' section for our definition of the chromosome 'cores'). Only two chromosomal inversions and two transpositions between chromosomes 9 and 11 were found, and two *E. cuniculi* genes were found to be duplicated and rearranged in *E. intestinalis* (an ABC transporter and a ubiquitin hydrolase). The average gene density of *E. intestinalis* (0.86 genes/kbp) was only fractionally higher than that

of *E. cuniculi* (0.84 genes/kbp) and the two genomes share the same reduced complement of introns. The *E. intestinalis* genes and introns were themselves reduced in size only slightly more than those of *E. cuniculi* (on an average, proteins are 0.06% shorter in *E. intestinalis*, Table 1), whereas the intergenic regions are reduced by 3.6% (Figs 1 and 2 and Table 1). Together with the slight reduction in the number of genes encoded within the chromosome 'cores' (1,819 of *E. intestinalis* as opposed to 1,830 in *E. cuniculi* when putative pseudogenes are not counted; Supplementary Data 2), reduction in the 'core' accounts for only about 34 kbp of the overall difference between the genome assemblies of *E. intestinalis* (95% assembled) and *E. cuniculi* (85% assembled), most of which is due to gene loss (Fig. 2).

In contrast to the cores, the *E. intestinalis* subtelomeres that we identified in this study are substantially reduced compared with those of *E. cuniculi*. Only three such regions have been completely characterized in *E. cuniculi* (two on chromosome 1, one on chromosome 4). We characterized all three corresponding regions in *E. intestinalis*, as well as two additional examples on chromosomes 5 and 10 (Fig. 3). These regions harbour a mixture of unique genes of known function in other organisms, hypothetical ORFs, and several members of a highly divergent *Encephalitozoon*-specific gene family (for example, DUF1609, DUF2463 or DUF1686) that seems to be actively transcribed[20,21]. In none of these cases is there any evidence that the coding regions represent pseudogenes, and there is similarly no indication that these regions contain any transposable elements that have been completely eradicated from both genomes. For the three chromosome ends known from *E. cuniculi*, the corresponding *E. intestinalis* chromosome is missing between 11 and 16 kbp of DNA and between 9 and 11 genes (Fig. 3). Moreover, in both cases in which the chromosome end is known in *E. intestinalis* but not in *E. cuniculi*, the latter can still be concluded to extend an additional 7 and 11 kbp, corresponding to five and nine genes, respectively (Fig. 3). This trend seems to extend across the entire genome: in all 22 chromosome ends, *E. intestinalis* is truncated relative to *E. cuniculi* for a total of 224 kbp, corresponding to 174 genes. Interestingly, the majority of these missing genes are not recognizable homologues of any known sequence in any other organism, except *E. cuniculi* (Fig. 2). Of these genes, 13 correspond to genes with homologues in other organisms, whereas 88 correspond to members of the DUF1609, DUF2463 or DUF1686 families, and 73 are hypothetical ORFs (Supplementary Data 2). Because our assembly is gap free but 108 kbp smaller than the estimated genome size, the remaining 708 kbp difference with the genome of *E. cuniculi* must be due to sequences at the ends of *E. cuniculi* chromosomes that are absent from *E. intestinalis*.

Conservation of non-coding sequences in *Encephalitozoon spp.* The significant imbalance between reduction at the chromosome ends and 'cores' suggests that the two regions are evolving under entirely different constraints. In particular, it suggests that the gene content and density of the chromosome 'cores' of both genomes have reached a certain limit. Deleterious effects of further reduction in gene content are relatively easy to imagine, as widespread gene loss presumably affects biochemical and cellular functions in an already unprecedentedly reduced proteome. The reason why the gene density has remained virtually unchanged is potentially related to the necessity of all remaining sequences to control expression. We examined this possibility by comparing sequence conservation across both genomes. Microsporidian genes are notorious for evolving very rapidly, but it is still surprising to find that the genes themselves displayed a much higher level of sequence divergence at synonymous sites than did the flanking intergenic spaces and introns (Fig. 4, Supplementary Table S2 and Supplementary Data 3). Importantly, there is no absolute correlation between the length of an intergenic space and its level of sequence conservation and, indeed, some of the longer regions are more conserved that the shortest ones (for example, the longest intergenic space in Fig. 4).

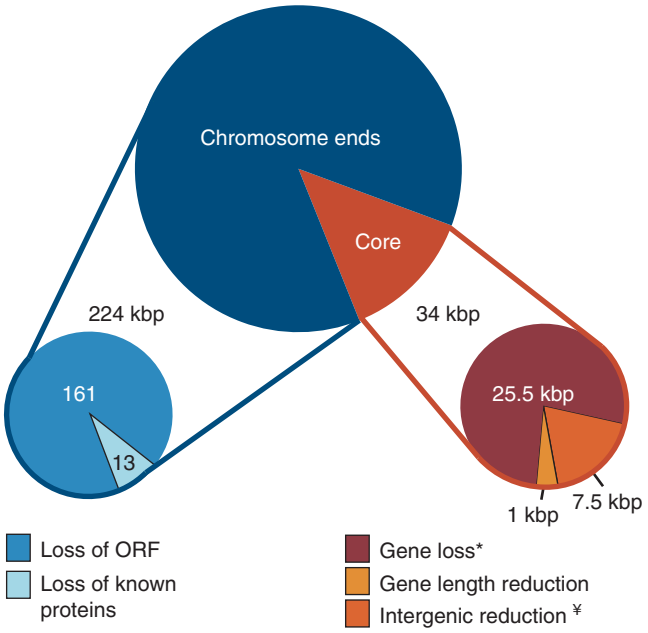

Figure 2 | Genome reduction in *E. intestinalis*. A comparison between the genome assembly of *E. intestinalis* (>95% assembled) with the available assembly of *E. cuniculi* (85% assembled) shows that the vast majority of genes that are absent in *E. intestinalis* are located at the chromosome ends in *E. cuniculi* (left, $n = 174$, 161 ORFs and 13 genes of identified function, Supplementary Data 2). At the chromosome cores, gene losses and shortening of intergenic regions account for most of the reduction in size of *E. intestinalis*, whereas reduction in gene length is negligible. * Includes their surrounding intergenic spaces. ¥Calculations are based on orthologous intergenic regions only.

Discussion

Although the *E. cuniculi* genome is among the smallest and most compact nuclear genomes known, the *E. intestinalis* genome is, at a still smaller 20%, at the known limit. By comparing the two, we can now determine which characteristics of a genome can be pushed to further extremes in these organisms, and which cannot. Although the majority of the size difference is due to gene loss, the protein-coding capacity of the two genomes is very similar because most of the genes that are absent in *E. intestinalis* are duplicates of genes that were retained, or unidentified ORFs. Indeed, the suite of unique genes in *E. intestinalis* is not only almost identical to that of *E. cuniculi*[3] but also scarcely different from any other known microsporidian[3,13,16–19]. The gene densities of *E. intestinalis* and *E. cuniculi* are also essentially the same, and the unusually high degree of conservation of intergenic sequences suggests that both genomes have also already been stripped down to a functional minimum, probably representing key regulatory elements located close to protein-coding genes in other organisms[22,23]. Overall, intergenic spaces and coding DNA seem to have followed similar evolutionary trajectories in maintaining what is essential for their function and getting rid of everything that is not.

The contrast between these similarities and the chromosome ends is stark, but is in line with previous findings based on pulsed field gel electrophoresis and other methodologies[20,24,25]. First, genes encoded near the ends of chromosomes are known to be evolving under different pressures than are other genes[26,27], and in *Encephalitozoon*, this is reflected by an even more elevated amino-acid sequence divergence compared with any other region of the

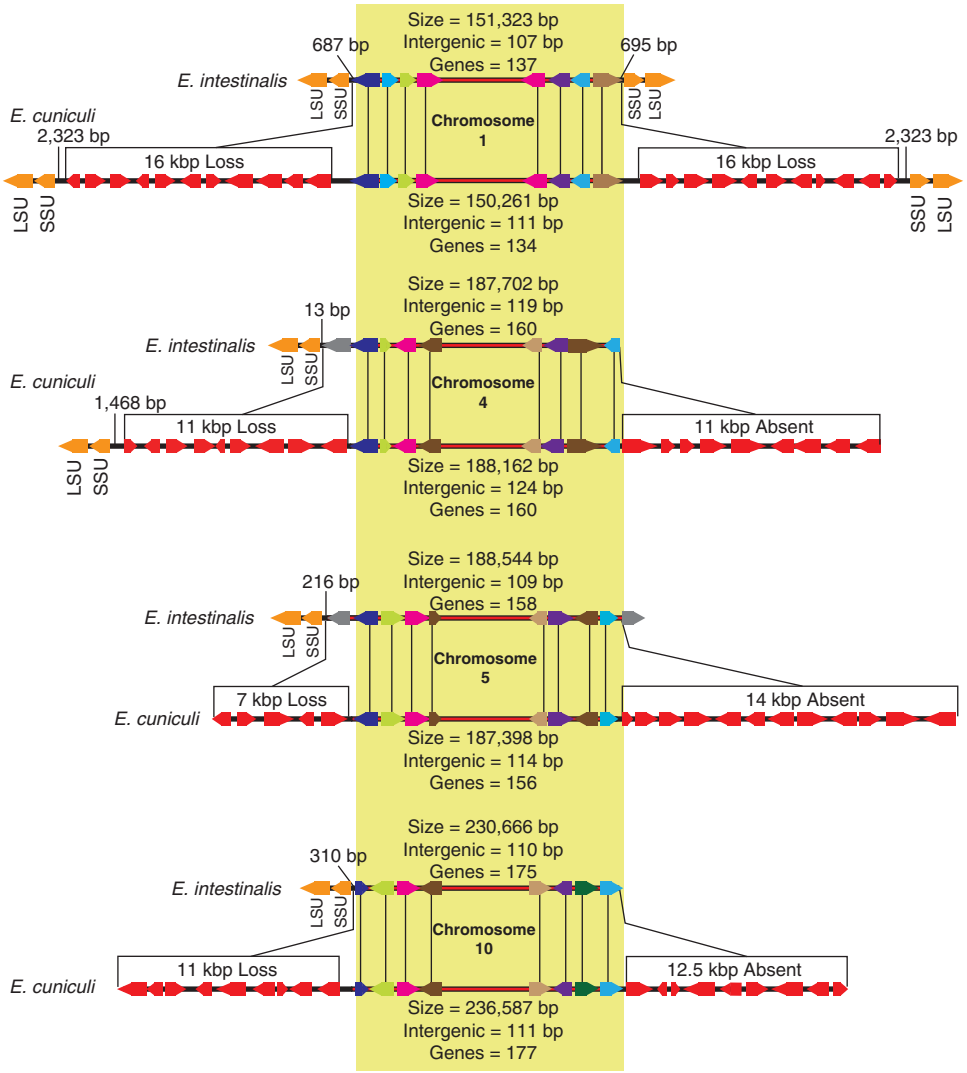

Figure 3 | Chromosome ends reduction in *E. intestinalis*. Structural comparisons between chromosomes 1, 4, 5 and 10 of *E. intestinalis* and *E. cuniculi*. Genes and their transcriptional direction are represented by rectangular arrows. The rDNA operon linked to chromosome ends is shown in orange. Genes that are absent from the *E. intestinalis* assembly are shown in red, whereas genes present in *E. intestinalis* but absent from *E. cuniculi* are grey. The chromosome 'cores' (shaded in a yellow box) contain long blocks of absolute colinearity between the two genomes: only the first and last four orthologues in these 'cores' are shown (as coloured boxes with directional arrows) for convenience. The total size of each 'core' is indicated for both species, along with the average length of its intergenic regions and the number of genes (including tRNAs and 5S rDNAs). SSU, small subunit ribosomal RNA gene; LSU, large subunit ribosomal RNA gene.

genome. Second, these regions are known to be highly plastic in *Encephalitozoon* species, and characterized by frequent gene losses and gains, even between strains of one species[24]. This massive difference in coding in the subtelomeric regions does, however, raise a number of interesting questions. It is noteworthy that we do not know whether *E. cuniculi* or *E. intestinalis* better represents their common ancestral condition: it is possible that a massive expansion has taken place at the ends of *E. cuniculi* chromosomes or that a reduction has taken place in the *E. intestinalis* lineage. In either cases, the high dispensability and rapid evolution of the hypothetical genes and gene families located in those regions raise the possibility that some of these genes in extant genomes may not be functionally significant. Although there is no direct evidence that any of these genes are non-functional (and many are known to be expressed[20]), it is possible that some of them either represent pseudogenes in the process of eroding or even spurious ORFs in the case of hypothetical ORFs. For the genes that do encode functional proteins, if the clustering of dispensable genes at chromosome ends is favoured because of the unusual conditions faced by the genes encoded there, then, paradoxically, the outright loss of these regions as in *E. intestinalis* would expose what were formerly 'core' genes to those conditions. Moreover, if these genes are expendable, why are they retained or even expanded in the otherwise severely compacted and reduced genome of *E. cuniculi*? It is also possible or even likely that our sampling of subtelomeric region may represent only a subset of the length diversity of these highly evolving genomic regions, as the genomic template used in this study comes from a pool of over 500 million spores. We should not rule out the possibility that no single sequence adequately represents any one chromosome end of a given species because of extensive subtelomere length variation within natural populations. The identification of additional subtelomeric regions in other, intermediate-sized *Encephalitozoon* species and between strains of each species will provide interesting insights into the presence of potential variation at these unusually evolving genomic regions, and also a miniature view of how different pressures shape different regions of nuclear genomes in general.

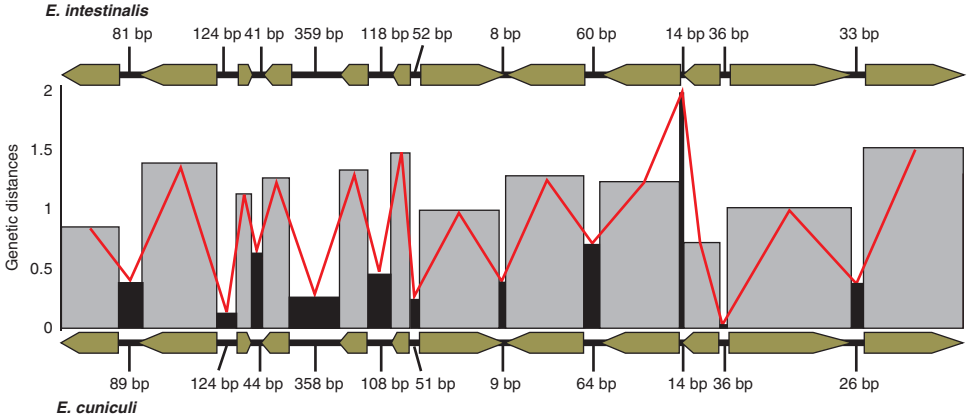

Figure 4 | Conservation of non-coding regions in *Encephalitozoon*. Schematic representation to scale of 12 orthologous genes (Ecu01_1080 to Ecu01_1170) and their intergenic spaces located on chromosome 1 of *E. intestinalis* (top) and *E. cuniculi* (bottom). Genes and orientations are represented by rectangular olive green arrows; intergenic regions are shown in black. The scale (left) represents the genetic distance between intergenic spaces and genes (number of substitutions per synonymous site for genes, or in total for intergenic regions). Grey rectangles represent the genetic distance between genes, whereas black rectangles represent the genetic distance between intergenic regions, the progression of which is tracked by the red line. With a single exception (the 14 bp intergenic region), the intergenic spaces are always more highly conserved than the genes they surround.

Methods

Cultivation and collection of *E. intestinalis* material. Spores from *E. intestinalis* (ATCC 50506; originally isolated from human alveolar lavage) were grown in the rabbit kidney fibroblast cell line, RK 13 (ATCC CCL-37), with RPMI 1640 supplemented with 5% fetal bovine serum, 2 mM L-glutamine and antibiotics (100 U penicillin per ml, 100 µg streptomycin per ml). T75 flasks were incubated at 37 °C with 5% CO_2, and culture medium was replaced two or three times per week. 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, TBS-Tween 20 (0.3%) and Tris buffered saline (TBS). The final pellet was then resuspended in TBS and mixed with an equal volume of 100% Percoll, followed by centrifugation at 400 *g* for 45 min at 4 °C. Host cell debris in the top 75% volume of Percoll was removed. The lower 25% volume of Percoll and the pellet were then transferred to a new tube, resuspended in TBS and washed several times. Owing to continued adherence of host cell (that is, rabbit) nucleic acid onto the spores, an additional series of washes were performed with TBS-SDS (0.1%), followed by three washes with TBS.

DNA extraction procedure. Genomic DNA was extracted from approximately 500 million spores. Spores were pelleted by centrifugation, resuspended in 300 µl of lysis solution (EPICENTRE Biotechnologies) containing Proteinase K and mixed thoroughly using a vortex. Glass beads (200 µl, 150–212 µm in diameter) were added to the sample, which was immediately incubated at 65 °C for 15 min and bead-beaten at 2,500 r.p.m. for 30 s every 5 min. The sample was then cooled to 37 °C and incubated for 30 min at the same temperature on addition of 2 µl of 5 µg µl^{-1} RNase A. After treatment with RNAse, the sample was placed on ice for 5 min, 150 µl of MPC Protein Precipitation Reagent (EPICENTRE Biotechnologies) was added and the solution was vortexed vigorously for 10 s. Protein debris was pelleted at 4 °C for 10 min at a speed of ≥10,000 *g* and the supernatant was transferred to a clean microcentrifuge tube. DNA was then precipitated using isopropanol, rinsed twice using 70% ethanol and the DNA was finally suspended in TE buffer.

Genome sequencing and *de novo* assembly. For deep sequencing, a genomic shotgun library was prepared from less than 200 ng of genomic DNA. We used Fasteris-modified bar-coded adapters to permit multiplexing, in this case using an ACTGT bar code at the beginning of the forward and reverse reads. Fragments with inserts of approximately 320–340 bp were selected, and an aliquot of the library was cloned into a TOPO plasmid and seven clones were sequenced by Sanger to confirm for insert sizes and that the constructs were derived from *Encephalitozoon*. The library was then subjected to deep sequencing using half a channel of the Illumina GAIIx instrument and 31 bp paired-end reads (with an average insert of 337 bp), resulting in 428,131,080 bp of unique DNA sequence. Reads were assembled using Velvet[29] with a hash value of 19, resulting in 137 scaffolds with an average size of 16,181 bp and an average coverage of 40×.

Genome completion and annotation and analysis. PCR using nested primers, cloning and Sanger sequencing were used in combination with Consed[29] to validate the breaks within the scaffolds and to link scaffold together into 11 continuous sequences homologous to the 11 chromosomes of *E. cuniculi*. Subtelomeric regions were obtained using nested PCR and Sanger sequencing, with forward primers designed on large subunit ribosomal RNA of *E. intestinalis* in combination with primers designed on the first and last identifiable genes from the 'core' of each chromosome. We define the chromosomal 'cores' heuristically as the genomic sequence located between the first and last recognizable *E. intestinalis* gene that has a clearly recognizable orthologue in *E. cuniculi*. Other chromosomal locations are referred to as 'subtelomeric' regions. Annotation was manually performed using Artemis[30] in combination with BLAST procedures[31] against the genome of *E. cuniculi* and the 'non-redundant' repository at NCBI. Transfer RNAs were detected using tRNAscan-SE[32] and spliceosomal introns were manually detected. Pairwise genetic distances between orthologous intergenic sequences of *E. intestinalis* and *E. cuniculi* were calculated using the maximum likelihood methodology implemented in MEGA4.[33] Pairwise genetic distances between fourfold degenerate (neutral) sites of orthologous genes were calculated using the Kumar method.[33] Sequences of the 11 individual *E. intestinalis* chromosomes are available from GenBank under accession numbers CP001942, CP001943, CP001944, CP001945, CP001946, CP001947, CP001948, CP001949, CP001950, CP001951 and CP001952.

References

1. Corradi, N. & Keeling, P. J. Microsporidia: a journey through radical taxonomical revisions. *Fungal Biol. Rev.* **23,** 1–8 (2009).
2. James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).
3. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
4. Lee, S. C. *et al.* Microsporidia evolved from ancestral sexual fungi. *Curr. Biol.* **18,** 1675–1679 (2008).
5. Van de Peer, Y., Ben Ali, A. & Meyer, A. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* **246,** 1–8 (2000).
6. Weiss, L. M. Microsporidia: emerging pathogenic protists. *Acta. Trop.* **78,** 89–102 (2001).
7. Keeling, P. J. *et al.* The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. *Genome Biol. Evol.* **2,** 304–309 (2010).
8. Canning, E. U. Nuclear division and chromosome cycle in microsporidia. *Biosystems* **21,** 333–340 (1988).
9. Corradi, N., Burri, L. & Keeling, P. J. mRNA processing in *Antonospora locustae* spores. *Mol. Genet. Genomics* **280,** 565–574 (2008).
10. Corradi, N., Gangaeva, A. & Keeling, P. J. Comparative profiling of overlapping transcription in the compacted genomes of microsporidia *Antonospora locustae* and *Encephalitozoon cuniculi*. *Genomics* **91,** 388–393 (2008).
11. Williams, B. A., Slamovits, C. H., Patron, N. J., Fast, N. M. & Keeling, P. J. A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl Acad. Sci. USA* **102,** 10936–10941 (2005).
12. Corradi, N. *et al.* Patterns of genome evolution among the microsporidian parasites *Encephalitozoon cuniculi*, *Antonospora locustae* and *Enterocytozoon bieneusi*. *PLoS ONE* **2,** e1277 (2007).
13. Slamovits, C. H., Fast, N. M., Law, J. S. & Keeling, P. J. Genome compaction and stability in microsporidian intracellular parasites. *Curr. Biol.* **14,** 891–896 (2004).

14. Peyretaillade, E. *et al.* Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucleic Acids Res.* **26,** 3513–3520 (1998).
15. Peyret, P. *et al.* Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora). *Genome Res.* **11,** 198–207 (2001).
16. Akiyoshi, D. E. *et al.* Genomic survey of the non-cultivatable opportunistic human pathogen, *Enterocytozoon bieneusi*. *PLoS Pathog.* **5,** e1000261 (2009).
17. Cornman, R. S. *et al.* Genomic analyses of the microsporidian *Nosema ceranae*, an emergent pathogen of honey bees. *PLoS Pathog.* **5,** e1000466 (2009).
18. Corradi, N., Haag, K. L., Pombert, J. F., Ebert, D. & Keeling, P. J. Draft genome sequence of the *Daphnia* pathogen *Octosporea bayeri*: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. *Genome Biol.* **10,** R106 (2009).
19. Williams, B. A. *et al.* Genome sequence surveys of *Brachiola algerae* and *Edhazardia aedis* reveal microsporidia with low gene densities. *BMC Genomics* **9,** 200 (2008).
20. Dia, N. *et al.* InterB multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species. *Curr. Genet.* **51,** 171–186 (2007).
21. Waters, P. F., Snowden, K. F. & Holman, P. J. A comparison of homologous genes encoding aminopeptidases among bird and human *Encephalitozoon hellem* isolates and a rabbit *E. cuniculi* isolate. *Parasitol Res.* **93,** 410–418 (2004).
22. Keightley, P. D. & Gaffney, D. J. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA* **100,** 13402–13406 (2003).
23. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3,** e42 (2005).
24. Brugere, J. F., Cornillot, E., Metenier, G. & Vivares, C. P. Occurence of subtelomeric rearrangements in the genome of the microsporidian parasite *Encephalitozoon cuniculi*, as revealed by a new fingerprinting procedure based on two-dimensional pulsed field gel electrophoresis. *Electrophoresis* **21,** 2576–2581 (2000).
25. Haro, M., Del Aguila, C., Fenoy, S. & Henriques-Gil, N. Intraspecies genotype variability of the microsporidian parasite *Encephalitozoon hellem*. *J. Clin. Microbiol.* **41,** 4166–4171 (2003).
26. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5,** e310 (2007).
27. Perry, J. & Ashworth, A. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9,** 987–989 (1999).
28. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
29. Gordon, D. Viewing and editing assembled sequences using Consed. *Curr. Protoc. Bioinformatics* **11,** Unit 11.2 (2003).
30. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672–2676 (2008).
31. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
32. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
33. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol. Evol.* **24,** 1596–1599 (2007).

Acknowledgments

We thank Sylvia Doan, David Twa, James T. Harper, Magne Osteras, Loïc Baerlocher and Lisa Bowers for help with sequencing, genome assembly and handling of microsporidian material. This work was supported by a Canadian Institutes for Health Research grant to P.J.K. (MOP-42517) and funding from the USA National Institutes of Health (RR00164 and AI071778) to E.S.D. P.J.K. is a Fellow of the Canadian Institute for Advanced Research (CIFAR) and a Senior Scholar of the Michael Smith Foundation for Health Research. N.C. is a scholar of the CIFAR and, at the time of the study, a Senior Postdoctoral Fellow of the Swiss National Science Foundation (PA00P3_124166). J.F.P. is the recipient of the Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT)/Génome Québec Louis-Berlinguet Postdoctoral Fellowship.

Author contributions

N.C., J.F.P. and P.J.K. contributed to the sequencing, annotation and assembly presented in this work and in the writing of the paper. L.F. contributed to sequencing. E.S.D. cultured and purified the spore material necessary for sequencing.

Additional information

Supplementary Information accompanies this paper on http://www.nature.com/naturecommunications

Competing financial interests: N.C., J.F.P., E.S.D. and P.J.K. declare no competing interests. L.F. is employed by Fasteris S.A., a company providing DNA sequencing services.

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

How to cite this article: Corradi, N. *et al.* The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nat. Commun.* 1:77 doi: 10.1038/ncomms1082 (2010).