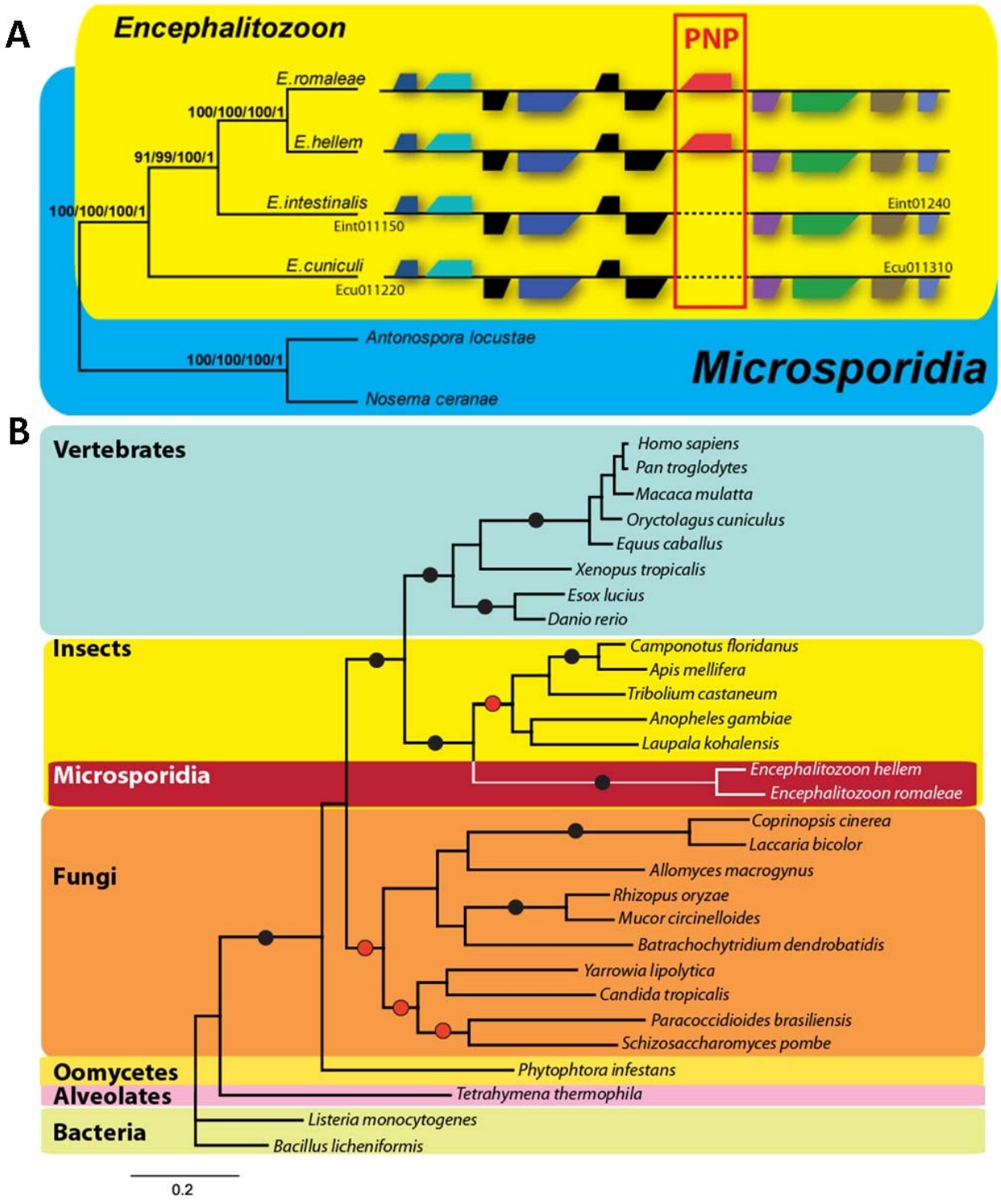
Supplemental Information

Acquisition of an animal gene by microsporidian intracellular parasites

Mohammed Selman, Jean-François Pombert , Leellen Solter , Laurent Farinelli , Louis M. Weiss, Patrick Keeling and Nicolas Corradi

Figure S1 (Related to Figure 1): A. Phylogeny of the genus *Encephalitozoon* and presence of a new gene in *Encephalitozoon romaleae* and *Encephalitozoon hellem.* B. Phylogenies of the PNP genes from *E. romaleae* and *E. hellem* after removal of the longest branches from Arthropods;

A. Alignment of highly collinear sections of approximately 13kb from *E. romaleae* and *E. hellem* that are homologous to regions of chromosome 1 of *Encephalitozoon cuniculi* and *E. intestinalis*, and phylogeny of the genus *Encephalitozoon* based on the amino acid sequence of 20 conserved genes (8155 amino acids in total). Values at nodes represent bootstrap and posterior probabilities support obtained using Minimum evolution (JTT model of evolution), as well as PhyML, RAxML, and Mr bayes (WAG model of evolution). The location of the purine nucleoside phosphorylase (PNP) genes in *E. romaleae* and *E. hellem* are shown in red. Conserved hypothetical proteins are shown as black rectangles, whereas genes encoding for proteins with known functions are shown as coloured rectangles. Dashed lines indicate portions of the genome sequence that are missing a particular gene. The GenBank accession number for the *E. romaleae* contig containing the PNP gene is JF808666. **B.** Phylogenetic relationships between the PNP genes of several eukaryotic and prokaryotic lineages based on 240 amino acid positions after removal of sequences corresponding to *Pediculus humanus* and Crustaceans. Major lineages are indicated by coloured boxes, while black circles indicate branches with bootstrap support of over 95% from Maximum Likelihood analyses (WAG model of evolution) and posterior probabilities over 0.95 obtained using Mr bayes (WAG model of evolution) and Phylobayes (CAT and LG models of evolution). Red circles indicate branches with posterior probabilities of 1 using Mr bayes, but with bootstrap support and posterior probabilities sometimes below 95% and 0.95 for either Maximum Likelihood analyses (WAG model of evolution), or for the Bayesian analyses performed under the CAT and LG models of evolution implemented in Phylobayes.


A
Encephalitozoon
PNP
E.romaleae
100/100/100/1
E.hellem
91/99/100/1
E.intestinalis
Eint011150
Eint01240
100/100/100/1
E.cuniculi
Ecu011220
Ecu011310
Antonospora locustae
100/100/100/1
Microsporidia
Nosema ceranae
B
Vertebrates
Homo sapiens
Pan troglodytes
Macaca mulatta
Oryctolagus cuniculus
Equus caballus
Xenopus tropicalis
Esox lucius
Danio rerio
Insects
Camponotus floridanus
Apis mellifera
Tribolium castaneum
Anopheles gambiae
Laupala kohalensis
Microsporidia
Encephalitozoon hellem
Encephalitozoon romaleae
Fungi
Coprinopsis cinerea
Laccaria bicolor
Allomyces macrogynus
Rhizopus oryzae
Mucor circinelloides
Batrachochytridium dendrobatidis
Yarrowia lipolytica
Candida tropicalis
Paracoccidioides brasiliensis
Schizosaccharomyces pombe
Oomycetes
Phytophtora infestans
Alveolates
Tetrahymena thermophila
Bacteria
Listeria monocytogenes
Bacillus licheniformis
0.2

Supplemental Experimental Procedures

Identification of putative animal-derived genes in the E. romaleae genome

Open reading frames (ORFs) were annotated across the genome survey of *Encephalitozoon romaleae* using Artemis [S1]. Potential LGTs of animal origin were searched, at first, across annotated ORFs using a number blast procedures (Blastp, Blastx, tBlastn, tBlastx) [S2] against available microsporidian genome data and the 'non-redundant' repository at NCBI. One gene found to be absent from any other available microsporidian sequence data (PNP), was also found to show strong sequence similarities against a number of metazoan sequences following Blast searches. Its metazoan origin was further tested using a number of phylogenetic methodologies, as explained in the section below**.**

Phylogenetic analyses

Amino acid sequences of PNP's eukaryotic orthologs (best reciprocal blast hits) were acquired from RefSeq GenBank, ESTdb, and from complete eukaryotic genome databases from the Broad institute and DOE Joint Genome Institute. GenBank accession numbers used for phylogenetic reconstruction can be found in the table below. Protein sequences were aligned using MUSCLE 3.7[S3], with a maximum number of iterations of 16. Poorly aligned positions and divergent regions of the alignment were removed using GBlocks 0.91b and default settings [S4], resulting in a data set of 174 amino acids. Trimal [S5], a less stringent trimming tool, was also used to remove non-informative amino acid regions from the alignment using the "strict method". The resulting alignment contained a suite of 240 amino acids.

Table: List of accession numbers (GenBank) and species used in phylogenetic analyses of PNP genes, Related to Figure 1.

PNP Accession	Species Name	Group
YP_079657	*Bacillus licheniformis ATCC 14580*	Bacteria
NP_465477	*Listeria monocytogenes EGD-e*	Bacteria
XP_001020972	*Tetrahymena thermophila*	Alveolata
EGF83310	*Batrachochytrium dendrobatidis JAM81*	Fungi
EEH43246	*Paracoccidioides brasiliensis Pb18*	Fungi
XP_506036	*Yarrowia lipolytica CLIB122*	Fungi
XP_002548301	*Candida tropicalis MYA-3404*	Fungi
NP_593927	*Schizosaccharomyces pombe 972h-*	Fungi
XP_001837849	*Coprinopsis cinerea okayama7#130*	Fungi
XP_001878763	*Laccaria bicolor S238N-H82*	Fungi
AMAG_14981.1**	*Allomyces macrogynus*	Fungi
RO3G_00999**	*Rhizopus oryzae*	Fungi
105288*	*Mucor circinelloides CBS277.49*	Fungi
EFX76980	*Daphnia pulex*	Crustacea
ACO14860	*Caligus clemensi*	Crustacea

XP_967070	*Tribolium castaneum*	Hexapoda
XP_001688760	*Anopheles gambiae str. PEST*	Hexapoda
XP_391850	*Apis mellifera*	Hexapoda
EFN71333	*Camponotus floridanus*	Hexapoda
XP_002427236	*Pediculus humanus corporis*	Hexapoda
EH641232		
EH634545***	*Laupala kohalensis*	Hexapoda
ACO14424	*Esox lucius*	Actinopterygii
NP_998476	*Danio rerio*	Actinopterygii
NP_001006720	*Xenopus (Silurana) tropicalis*	Amphibia
XP_002718082	*Oryctolagus cuniculus*	Mammalia
XP_001104622	*Macaca mulatta*	Mammalia
NP_000261	*Homo sapiens*	Mammalia
XP_001140576	*Pan troglodytes*	Mammalia
XP_001505186	*Equus caballus*	Mammalia
XP_002897613	*Phytophthora infestans T30-4*	Oomycetes

***** Protein ID, JGI (DOE Joint Genome Institute)
******Locus ID, Broad Institute
*** EST fragment containing PNP gene

Phylogenetic analyses were carried out using Maximum Likelihood and Bayesian methods. Maximum Likelihood phylogenies were performed with PhyML 3.0 [S6] using the WAG substitution model, 100 bootstraps, 4 substitution rate categories and estimated gamma (Γ) parameters and proportion of invariant sites (I). Bayesian reconstructions were performed with Mr Bayes 3.1.2 [S7] under the WAG+Γ4+I model of amino acid substitution. The Markov chain Monte Carlo searches were run for 10,000,000 generations, sampling the Markov chains every 10 generations; the first 25,000 trees were discarded as 'burn-in'. Finally, the CAT and LG models implemented in Phylobayes were also used to reconstruct the phylogeny of PNP [S8]. The two independent chains run for 10,000 (CAT) and 50,000 (LG) cycles, even though both were found to rapidly converge (after 100 cycles). The posterior distributions obtained under these independent runs were compared after a burn-in of 100 (CAT) and 1,000 (LG), resulting in maxdiff values much less than 0.1 (indicative of very good runs) for CAT and lower than 0.016 for LG (indicative of very good runs). The consensus trees were obtained by pooling all the trees from both chains.

The phylogeny of the genus *Encephalitozoon* (Figure S1.A) was reconstructed using the amino acid sequences of 20 conserved genes (8155 amino acids in total) from the complete genomes of *E. cuniculi* and *E. intestinalis*, the genome sequence survey data from *E. romaleae*, and sequences from an ongoing genome project from *E. hellem*. Homologues from complete or nearly complete genomes of *Nosema ceranae* and *Antonospora locustae* were used as outgroups .

Table: List of accession numbers (GenBank) and species used in phylogenetic analyses conserved Microsporidian genes.

Protein	*Antonospora locustae**	*Nosema ceranae*	*E. cuniculi*	*E. hellem*	*E. intestinalis*	*E. romaleae*
Actin	AAB86863	XP_002995436	XP_965880	AAB86862	XP_003072256	JN039386
DNA repair helicase RAD25	contig_340	XP_002995956	XP_965942	JN039409	XP_003072319	JN039394
Enolase	contig_489	XP_002995378	NP_586285	JN039405	XP_003073850	JN039388
Glucose-6-phosphate isomerase	contig_2954	XP_002996045	NP_597407	JN039408	XP_003072875	JN039392
Hsp70NP	AAC47660	XP_002995188	NP_586360	BAB69033	XP_003073899	JN039393
Isoleucyl tRNA synthetase	AAC41564	XP_002996071	CAD26020	BAD83624	XP_003073958	JN039397
Mannose-1-phosphate-guanylyltransferase	contig_1173	XP_002996275	NP_586375	JN039402	XP_003073916	JN039383
Methionine aminopeptidase 2	contig_868	XP_002996537	NP_586190	AAP51023	ADM12396	JN039385
MCM2	contig_119	XP_002995530	CAD25272	JN039410	ADM11370	JN039396
Pyruvate dehydrogenase E1 alpha subunit	contig_369	XP_002996136	XP_955659	JN039403	XP_003073591	JN039384
Phosphomannomutase	contig_1060	XP_002994973	NP_597365	JN039401	XP_003072832	JN039381
Pyruvate kinase	contig_493	XP_002996514	XP_955618	JN039404	XP_003073549	JN039387
RNA polymerase I largest subunit	AAT12325	XP_002996654	NP_584825	JN039411	XP_003073214	JN039398
Pyruvate dehydrogenase E1 beta subunit	contig_369	XP_002995305	NP_584800	JN039400	XP_003072763	JN039380
RNA polymerase II largest subunit	AAD12605	XP_002995402	CAD26175	JN039412	XP_003072524	JN039399
Trehalose-6-phosphate phosphatise	AAT12365	XP_002996623	XP_965922	JN039406	XP_003072299	JN039395
Translation elongation factor 1 alpha	contig_559	XP_002995330	NP_584794	JN039407	XP_003072757	JN039391
Tubulin alpha	AAC47419	XP_002995388	NP_586048	P92120	XP_003073238	JN039390
Tubulin beta	AAG48935	XP_002995929	NP_597591	Q24829	XP_003072575	JN039389
Transcription initiation factor TFIIB	contig_106	XP_002996560	NP_585866	JN052740	XP_003073070	JN039382

*Contigs from *A. locustae* can be found at http://forest.mbl.edu/cgi-bin/site/antonospora01.

Extended acknowledgements

We would also like to thank Sara Selman, Douglas Whitman, Magne Osteras and Loïc Baerlocher for help with sequencing, genome assembly and handling of microsporidian material, and Stephane Aris-Brosou, Toni Gabaldon and two anonymous reviewers for important comments on a previous version of this manuscript. We acknowledge access to the *Allomyces*

macrogynus and *Rhizopus oryzae* genome data generated by the Broad Institute and to the *Batrachochytrium dendrobatidis* and *Mucor circinelloides* genome data provided by the US Department of Energy Joint Genome Institute. We thank Hilary Morrison and acknowledge the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution for the use of data included in the *Antonospora locustae* Genome Project funded by NSF.

Supplemental References

S1. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics (Oxford, England) *16*, 944-945.

S2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

S3. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

S4. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

S5. Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England) *25*, 1972-1973.

S6. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*, 696-704.

S7. Huelsenbeck, J.P., and Ronquist, F. (2001). MR BAYES: Bayesian inference of phylogenetic trees. Bioinformatics (Oxford, England) *17*, 754-755.

S8. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics (Oxford, England) *25*, 2286-2288.