

**The following resources related to this article are available online at
www.sciencemag.org (this information is current as of December 6, 2009):**

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/309/5731/131>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/309/5731/131/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/309/5731/131#related-content>

This article **cites 18 articles**, 7 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/309/5731/131#otherarticles>

This article has been **cited by** 82 article(s) on the ISI Web of Science.

This article has been **cited by** 14 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/309/5731/131#otherarticles>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

Genome of the Host-Cell Transforming Parasite *Theileria annulata* Compared with *T. parva*

Arnab Pain,^{1*} Hubert Renaud,¹ Matthew Berriman,¹ Lee Murphy,¹ Corin A. Yeats,^{1†} William Weir,² Arnaud Kerhornou,¹ Martin Aslett,¹ Richard Bishop,³ Christiane Bouchier,⁴ Madeleine Cochet,⁵ Richard M. R. Coulson,⁶ Ann Cronin,¹ Etienne P. de Villiers,³ Audrey Fraser,¹ Nigel Fosker,¹ Malcolm Gardner,⁷ Arlette Goble,¹ Sam Griffiths-Jones,¹ David E. Harris,¹ Frank Katzer,⁸ Natasha Larke,¹ Angela Lord,¹ Pascal Maser,⁹ Sue McKellar,² Paul Mooney,¹ Fraser Morton,¹ Vishvanath Nene,⁷ Susan O'Neil,¹ Claire Price,¹ Michael A. Quail,¹ Ester Rabbino-witsch,¹ Neil D. Rawlings,¹ Simon Rutter,¹ David Saunders,¹ Kathy Seeger,¹ Trushar Shah,³ Robert Squares,¹ Steven Squares,¹ Adrian Tivey,¹ Alan R. Walker,⁸ John Woodward,¹ Dirk A. E. Dobbelaere,¹⁰ Gordon Langsley,⁵ Marie-Adele Rajandream,¹ Declan McKeever,^{8,11} Brian Shiels,² Andrew Tait,² Bart Barrell,¹ Neil Hall^{1‡}

Theileria annulata and *T. parva* are closely related protozoan parasites that cause lymphoproliferative diseases of cattle. We sequenced the genome of *T. annulata* and compared it with that of *T. parva* to understand the mechanisms underlying transformation and tropism. Despite high conservation of gene sequences and synteny, the analysis reveals unequally expanded gene families and species-specific genes. We also identify divergent families of putative secreted polypeptides that may reduce immune recognition, candidate regulators of host-cell transformation, and a *Theileria*-specific protein domain [frequently associated in *Theileria* (FAINT)] present in a large number of secreted proteins.

Theileria are the only intracellular eukaryotic pathogens capable of reversibly transforming their host cells. *Theileria annulata* (TA) and *T. parva* (TP) are tick-borne hemoparasites (1) that give rise to lymphoproliferative diseases (2) of cattle known, respectively, as tropical theileriosis and East Coast fever (ECF). The molecular mechanisms are unknown, but previous analyses indicate that both species subvert the same host-cell signal transduction pathways (3). Although the parasites have similar life cycles involving intracellular stages in leukocytes and in red blood cells, they are transmitted by different tick species and transform different cell types. In contrast to ECF, cases of tropical theileriosis are accompanied by severe anemia. Available therapeutics are reliable only in the early stages of disease, and existing vaccines rely on the administration of live parasites. There is an urgent need for improved control and therapeutics.

The nuclear genome (4) of TA is similar in size (8.35 Mb) to that of TP (8.3 Mb); it spans four chromosomes that range from 1.9 to 2.6 Mb (Table 1 and table S1). We predicted 3792 putative protein-coding genes in TA. In addition, a total of 49 tRNA and 5 ribosomal RNA (rRNA) genes were found, revealing common features in rRNA units

between the species (5) (table S1). The telomeres and presumptive centromeres of TA and TP are similar in base composition, size, and arrangement.

Like many parasitic protozoa, both *Theileria* spp. have tandem arrays of genus-specific, hypervariable gene families (6) (table S3) that map adjacent to the telomeres (6) with an overall arrangement that appears conserved (Fig. 1). Most of these subtelomeric genes encode predicted secreted proteins. Genes previously described as related to the restric-

Table 1. Comparison of protein coding genes in *T. annulata* and *T. parva*. Unique genes are calculated by filtering the genes without orthologs; members of gene families with counterparts in both genomes are removed, as are any that have a translated query versus translated database (TBLASTX) hit in the other species (e value $< 1 \times 10^{-10}$). Genes smaller than 100 amino acids were manually checked.

	<i>T. annulata</i>	<i>T. parva</i>
Genome size	8351610	8308027
G+C content	32.54	34.1
Gene number	3792	4035
Genes with orthologs	3265	3265
Genes without orthologs	493	710
Unique genes	34	60

tion enzyme *Sfi*I fragment (designated family 3, table S3) are found proximal to the telomeres (Fig. 1B), followed by Pro/Gln-rich proteins (family 1, table S3). The boundary between subtelomeric gene families and “housekeeping” genes is defined by adenosine 5'-triphosphate-binding cassette (ABC) transporter genes (family 5, table S3) in the opposite coding orientation. Stage-specific expressed sequence tags (ESTs) indicate that at least three subtelomeric ABC transporters are constitutively transcribed in macroschizont, merozoite, and piroplasm stages in the mammalian host. Members of gene families 3 and 5 also occur internally in the genome. Our findings are consistent with vigorous genetic exchange between subtelomeres, fostering expansion and diversification of antigens, with internal clusters that may act as reservoirs.

The nonsubtelomeric regions of the TA and TP genomes show strong conservation of synteny with only a few inversions of small sequence blocks and no interchromosomal rearrangements (Fig. 1A). Short interruptions to synteny corresponded to the insertion or deletion of genes, and usually involve members of large gene families, as exemplified by the TP repeat (*Tpr*) genes (4) and their *Tpr*-related counterparts in TA (*Tar*). These *Tar* genes form the second largest family in both genomes. The majority of *Tpr* genes form a single array on TP chromosome 3 (5, 7), located at a large inversion point. *Tar* genes are dispersed throughout the four chromosomes in TA and cause small interruptions in synteny. The lower sequence divergence between *Tpr* compared with *Tar* genes suggests

¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

²Division of Veterinary Infection and Immunity, Parasitology Group, Institute of Comparative Medicine, Faculty of Veterinary Medicine, Bearsden Road, Glasgow G61 1QH, UK. ³The International Livestock Research Institute (ILRI), Post Office Box 30709, Nairobi, Kenya.

⁴Plate-Forme Génomique-Pasteur Génomique, Ile de France Institut Pasteur, 25–28 rue du Docteur Roux, 75724 Paris, France. ⁵Unité de Recherche Associée CNRS 2581, Département de Parasitologie, Bâtiment Elie Metchnikoff, Institut Pasteur, 25–28 rue du Docteur Roux, 75724 Paris Cedex 15, France. ⁶European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ⁷The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. ⁸Division of Veterinary Clinical Studies, Royal School of Veterinary Studies, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, UK. ⁹Institute of Cell Biology, University of Bern, Baltzerstrasse 4, 3012 Bern, Switzerland. ¹⁰Molecular Pathology, Institute of Animal Pathology, University of Bern, Laenggassstrasse 122, 3012 Bern, Switzerland. ¹¹Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 OPZ, UK.

¹⁰Unité de Recherche Associée CNRS 2581, Département de Parasitologie, Bâtiment Elie Metchnikoff, Institut Pasteur, 25–28 rue du Docteur Roux, 75724 Paris Cedex 15, France. ⁶European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ⁷The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. ⁸Division of Veterinary Clinical Studies, Royal School of Veterinary Studies, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, UK. ⁹Institute of Cell Biology, University of Bern, Baltzerstrasse 4, 3012 Bern, Switzerland. ¹⁰Molecular Pathology, Institute of Animal Pathology, University of Bern, Laenggassstrasse 122, 3012 Bern, Switzerland. ¹¹Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 OPZ, UK.

¹¹Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 OPZ, UK.

*To whom correspondence should be addressed. E-mail: ap2@sanger.ac.uk

†Present address: Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK.

‡Present address: The Institute for Genomic Research, Rockville, MD 20850, USA.

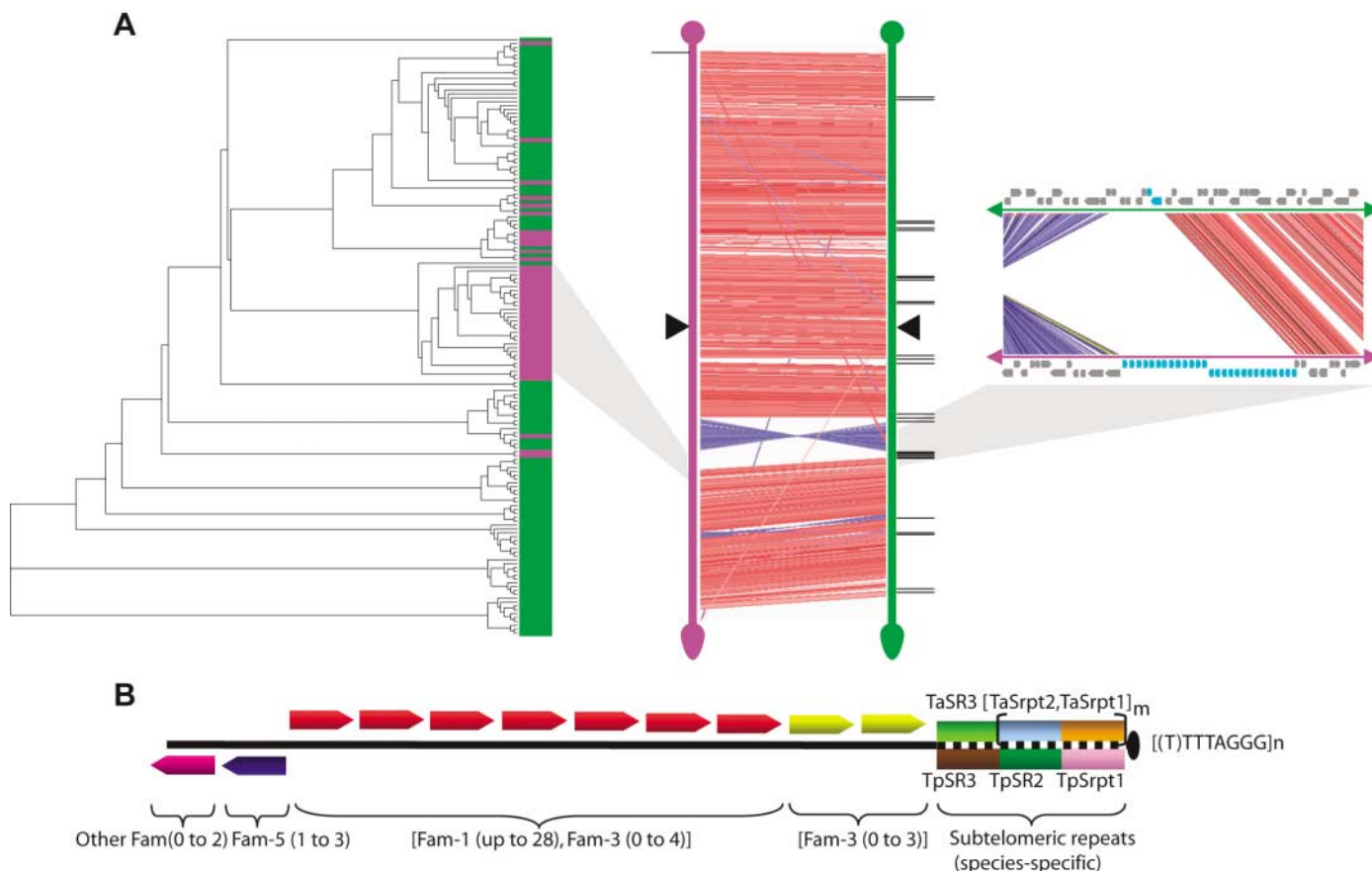


Fig. 1. Large-scale synteny between *T. annulata* and *T. parva* chromosomes. (A) Synteny breaks of chromosome 3 of TA (green) and TP (purple) are located at *Tpr* genes. (Middle) Chromosome 3 of TA and chromosome 3 of TP are aligned. Connecting lines show maximal unique matches between the two chromosomes. Red lines, alignments in the same orientation; blue lines, alignments in opposing orientations; black triangles, putative centromeres; black lines, *Tpr* genes occurring outside the *Tpr* locus. The position of the *Tpr* locus of TP is aligned with the gray shaded area. (Left) The phylogenetic tree shows the clustering of the TP genes when compared with the TA genes. Branches ending in green boxes represent TA genes and purple boxes

represent TP genes. All genes in the *Tpr* locus occur in the cluster which is aligned with the gray shaded area. (Right) A close-up of the insertion of the *Tpr* locus in TP (purple) with respect to TA (green), with *Tpr* and *Tar* genes (blue) and all other genes (gray). (B) Organization of a representative subtelomere (not to scale). The black line represents the coding part of the subtelomere, with the arrangement of gene families (arrowheads) shared between TA and TP. The arrowheads indicate the transcriptional orientation; the observed range in numbers of genes is in parentheses. The dotted black line represents the species-specific noncoding regions (upper, TA; lower, TP). Srpts, subtelomeric repeats; SR, subtelomeric regions (4).

that they expanded after speciation. The single array in TP may allow gene conversion to prevent divergence.

Noncoding regions of subtelomeres are complex. In TA, from the terminus inward, a succession of paired guanine-cytosine (GC)-rich subtelomeric repeats (TaSrpt1 and TaSrpt2) are followed by a single-copy sequence at all chromosome ends (TaSR3; Fig. 1B and fig. S3). No such repeats are found in TP subtelomeres; a terminal sequence (TpSrpt1, ~140 base pairs) is shared by all chromosomal ends, followed by a thymine-rich region (TpSR2), then by a region shared by many but not all TP subtelomeres (TpSR3).

We predicted 3265 orthologous genes between the genomes. Most genes without orthologs are members of gene families; only a small proportion (34 in TA, 60 in TP; table S4) are single-copy genes to which functions could not be ascribed, but EST data detected that four of these are expressed in TA. No major species differences were found in the numbers

of predicted transcription-associated proteins, peptidases (4), or core metabolic enzymes (5).

We evaluated evolutionary pressure acting on genes using the ratio of nonsynonymous to synonymous substitutions (dN/dS) between orthologs (table S7). This method can potentially identify immunogenic genes and thus putative vaccine candidates (8). Where possible, we matched dN/dS with stage-specific expression patterns from the EST data in TA. Constitutively expressed genes displayed the lowest dN/dS values (Fig. 2). Similar to *Plasmodium* (9), genes encoding merozoite surface proteins yielded the highest dN/dS ratios (Fig. 2); these proteins are candidates for immune selection (10). For predicted macroschizont polypeptides with signal peptides, dN/dS values were also high, although lower than those for merozoites. Surprisingly, genes encoding macroschizont glycosylphosphatidylinositol (GPI)-anchored membrane proteins have dN/dS values similar to housekeeping genes. In contrast, high dN/dS ratios

were found for macroschizont proteins without predicted membrane retention motifs that are potentially secreted into the leukocyte cytosol. The high dN/dS values associated with host-exported *Theileria* proteins might reflect regulatory functions that have diversified after speciation of TA and TP. Alternatively, they might reflect exposure to the immune system, after rapid degradation to generate peptides presented by major histocompatibility complex antigens on the infected cell surface. Consistent with this, PEST (a signal for rapid proteolytic degradation) regions (11) were identified in many of these polypeptides (table S8).

Almost all members of the major *Theileria*-specific subtelomeric protein family members incorporate varying numbers (1 to 54) of a single, highly polymorphic domain with an average length of 70 residues, a designation frequently associated in *Theileria* (FAINT), formerly known as DUF529 (12). Over 900 copies were found in 166 TA proteins and in equivalent numbers of TP proteins (fig. S5).

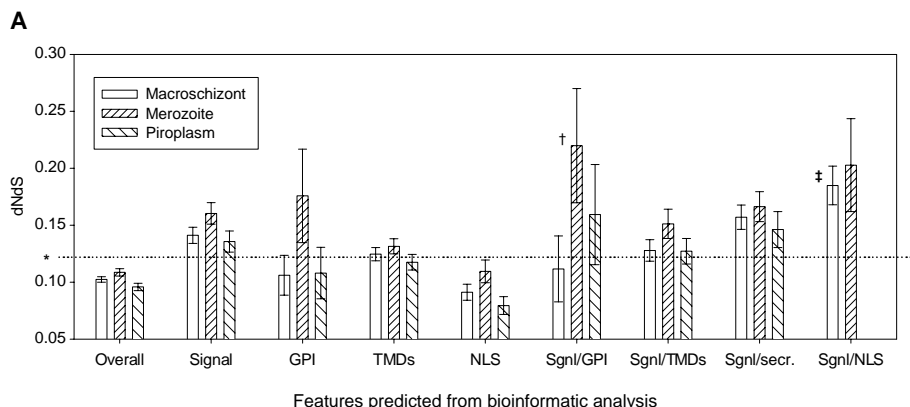


Fig. 2. (A) dN/dS ratios computed between pairs of orthologous genes in TA and TP. Mean dN/dS values of expressed proteins as a function of life-cycle stage in TA and predicted protein motifs and signals. Error bars show means \pm SE. EST data were from cDNAs from three life-cycle stages in TA (macroshizont, merozoite, and piroplasm). Grouping of proteins was based on presence of certain domains (4), indicated as follows: Signal, presence of a signal peptide; GPI, GPI anchor; TMD, transmembrane domain; NLS, nuclear localization sequence; secr., secreted. We assume where GPIs occurred in the absence of signal peptides, it was

The majority of the FAINT domain-containing proteins have no other recognizable domains except a putative signal peptide, consistent with export to the host. However, in members of the TashAT gene cluster, one or more FAINT domains appear with AT-hook and PEST motifs on the same protein (13, 14) (fig. S5 and table S8). We found only one other FAINT domain containing protein in the UniProt protein database (15), occurring in a nontransforming *Theileria* (synonym of *Babesia equi*), which also invades leukocytes and develops to a macroshizont stage (16). We also described proteins containing previously unrecognized short amino acid repeat domains in both genomes (4). The species-specific nature of the domains suggests that they have expanded recently (4) (fig. S1).

The parasite genes involved in host-cell transformation must be expressed by the macroshizont stage, and their products must be released into the host cell cytoplasm or expressed on the parasite surface. This would generally require a signal peptide or a specific host-targeting signal sequence. Candidates meeting these criteria include the previously described TashAT and SuAT protein families in TA (13, 14) and related TP host nuclear proteins (TpHNs) in TP. In addition to localizing to the host nucleus, members of the TashAT family bear cyclin-dependent kinase phosphorylation motifs, cyclin docking sites, and AT-hook DNA binding domains (table S8). A cluster of 17 SuAT1- and TashAT-like genes was identified in the TA genome and an orthologous gene family of 20 members in a syntenic region of the TP genome. However, TpHNs lack consensus AT-hook motif, a divergence that could be interpreted as a result of species adaptation to their preferred host-cell type.

We screened both predicted proteomes with a database of proteins linked to cell transformation and tumor progression (17) and matched the hits with the presence of a signal peptide and the macroshizont EST data set (4). No obvious proto-oncogenes, kinases, or phosphatases were identified. However, this screen did identify members of the HSP90 subfamily, DEAD-box RNA helicases, peptidases, immunophilins, members of the thioredoxin/glutaredoxin family, and leucine-zipper proteins (table S9).

Proteins that function in lipid metabolism were also identified as transformation candidates. First, we found proteins related to phospholipase A2, whose activity is elevated in tumor cells (18), in both predicted proteomes and, unlike in other apicomplexan parasites, they carry a signal peptide. Second, choline kinase genes (ChoKs) are present at high copy number compared with other apicomplexans. ChoK activity is deregulated in transformed cell lines and its inhibition results in a reversible blockage of cell proliferation (19). Finally, the cell cycle effectors uridine phosphorylases and leucine carboxyl methyltransferases (20), whose activity is raised in tumor cells (21), are tandemly duplicated in TA and TP. However, no signal sequence is predicted for the latter three enzymes, so it remains to be determined whether their expansion reflects the ability of the macroshizont to maintain host-cell transformation.

References and Notes

1. M. T. Allsopp, T. Cavalier-Smith, D. T. De Waal, B. A. Allsopp, *Parasitology* **108**, 147 (1994).
2. L. M. Forsyth *et al.*, *J. Comp. Pathol.* **120**, 39 (1999).
3. D. A. Dobbelaere, P. Kuenzi, *Curr. Opin. Immunol.* **16**, 524 (2004).
4. Materials and methods are available as supporting material on Science Online.

B

Protein-coding regions	
Av % protein ID	83.7
Av % nucleotide ID	82.4
Av dN	0.09
Av dS	0.82
Av dN/dS	0.097
Non-protein coding regions	
% nucleotide ID	74

because of the limitations of gene boundaries and in the prediction software. Dotted line marked by asterisk, 0.1220, average dN/dS across all genes with orthologs; †, merozoite/signal/GPI proteins versus other merozoite proteins ($P = 0.016$; 95% CI: 0.0214 to 0.2080), Mann-Whitney test; ‡, macroshizont/signal/NLS proteins versus other macroshizont proteins ($P = 0.001$; 95% CI: 0.04831 to 0.13320), Mann-Whitney test. (B) Summary of the analysis. The average (Av) dN/dS ratios and identities (ID) of coding and noncoding regions are shown for all orthologous genes between TA and TP.

5. M. J. Gardner *et al.*, *Science* **309**, 134 (2005).
6. J. D. Barry, M. L. Ginger, P. Burton, R. McCulloch, *Int. J. Parasitol.* **33**, 29 (2003).
7. H. A. Baylis, S. K. Sohal, M. Carrington, R. P. Bishop, B. A. Allsopp, *Mol. Biochem. Parasitol.* **49**, 133 (1991).
8. T. Endo, K. Ikeo, T. Gojoberi, *Mol. Biol. Evol.* **13**, 685 (1996).
9. N. Hall *et al.*, *Science* **307**, 82 (2005).
10. M. J. Gubbels, F. Katzer, B. R. Shiels, F. Jongejan, *Parasitology* **123**, 553 (2001).
11. M. Rechsteiner, S. W. Rogers, *Trends Biochem. Sci.* **21**, 267 (1996).
12. A. Bateman *et al.*, *Nucleic Acids Res.* **32**, D138 (2004).
13. D. G. Swan, K. Phillips, A. Tait, B. R. Shiels, *Mol. Biochem. Parasitol.* **101**, 117 (1999).
14. B. R. Shiels *et al.*, *Eukaryot. Cell* **3**, 495 (2004).
15. R. Apweiler *et al.*, *Nucleic Acids Res.* **32**, D115 (2004).
16. H. Mehlhorn, E. Schein, *Parasitol. Res.* **84**, 467 (1998).
17. More information about the cancer-related protein database is available at www.cancerindex.org/geneweb/.
18. P. Sved *et al.*, *Cancer Res.* **64**, 6934 (2004).
19. A. Ramirez de Molina *et al.*, *Oncogene* **21**, 4317 (2002).
20. T. Tolstykh, J. Lee, S. Vafai, J. B. Stock, *EMBO J.* **19**, 5682 (2000).
21. A. Kanzaki *et al.*, *Int. J. Cancer* **97**, 631 (2002).
22. We acknowledge the support of the Wellcome Trust Sanger Institute core sequencing and informatics groups. We thank N. Zidane and S. Duthoy for sequencing the macroshizont ESTs and V. Heussler and I. Roditi for helpful advice with this manuscript. The sequence and annotation of *T. annulata* genome have been submitted to the EMBL databases with consecutive accession numbers between CR940346 and CR940353; they can be viewed at www.GenBank.org. The EST sequences from all three life-cycle stages have been submitted to the EMBL database with consecutive accession numbers between AJ920420 and AJ936931. This work was supported by the Wellcome Trust.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5731/131/DC1
 Materials and Methods
 Figs. S1 to S5
 Tables S1 to S9
 References

31 January 2005; accepted 5 May 2005
 10.1126/science.1110418