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

Arnab Pain,1* Hubert Renault,1 Matthew Berriman,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,7 Audrey Fraser,1 Nigel Fosker,1 Malcolm Gardner,7 Arlette Goble,1 Sam Griffiths-Jones,7 David E. Harris,1 Frank Katzer,8 Natasha Larke,1 Angela Lord,1 Pascal Maser,9 Sue McKeever,2 Paul Mooney,1 Fraser Morton,1 Vishvanath Nene,7 Susan O’Neill,1 Claire Price,1 Michael A. Quail,1 Ester Rabbinowitsch,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. Dobbelelaere,10 Gordon Langsley,5 Marie-Adele Rajandream,1 Declan McKeever,8,11 Brian Shiels,2 Andrew Tait,2 Bart Barrett,1 Neil Hall1†

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 unequaly 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 anaemia. 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 putative secreted proteins. Genes previously described as related to the restriction enzyme SfiI 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 malarial 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 (3, 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 suggests

Table 1. Comparison of protein coding genes in T. annulata and T. parva. Unique genes are calculated by filtering the genes with 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 × 10^-10). Genes smaller than 100 amino acids were manually checked.

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


The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. Division of Veterinary Infection and Immunity, Parasitology Group, Institute of Comparative Medicine, Faculty of Veterinary Medicine, Bearsden Road, Glasgow G61 1QH, UK. The International Livestock Research Institute (ILRI), Post Office Box 30709, Nairobi, Kenya. Unité de Recherche Associée CNRS 2581, Département de Parasitologie, Bâtiment Elie Metchnikoff, Institut Pasteur, 25–28 rue du Docteur Roux, 75724 Paris, France. Unité de Recherche Associée CNRS 2581, Département de Parasitologie, Bâtiment Elie Metchnikoff, Institut Pasteur, 25–28 rue du Docteur Roux, 75724 Paris, France. European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. Division of Veterinary Clinical Studies, Faculty of Veterinary Medicine, University of Cambridge, Cambridge CB10 1SD, UK. MalariaPathology, Institute of Animal Pathology, University of Bern, Laenggasse 122, 3012 Bern, Switzerland. Molecular Pathology, Institute of Animal Pathology, University of Bern, Laenggasse 122, 3012 Bern, Switzerland. Morehead Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 0PZ, 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.
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, Ê140 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); 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).
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 macroschizont 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 macroschizont 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 (TpHINs) 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, TpHINs 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 macroschizont 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 phosphoribosyltransferases 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 macroschizont to maintain host-cell transformation.

References and Notes
4. Materials and methods are available as supporting material on Science Online.
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).
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 macroschizont ESTs and V. Heusler 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.GeneDB.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
Editor's Summary

This copy is for your personal, non-commercial use only.

Article Tools
Visit the online version of this article to access the personalization and article tools:
http://science.sciencemag.org/content/309/5731/131

Permissions
Obtain information about reproducing this article:
http://www.sciencemag.org/about/permissions.dtl

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

Science 309 (5731), 131-133. [doi: 10.1126/science.1110418]