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.

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.

<table>
<thead>
<tr>
<th>Gene Feature</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>GC 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>4933</td>
<td>7100</td>
</tr>
<tr>
<td>Unique genes</td>
<td>34</td>
<td>60</td>
</tr>
</tbody>
</table>
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), the observed range in numbers of genes is in parentheses. The dotted black line represents the species-specific noncoding regions (upper, TA; lower, TP).

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 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).

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. Candidate meeting these criteria include the previously described TashAT and SuAT protein families in TA (13, 14) and related TP host nuclear proteins (TphHns) 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, TPHs 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 apicomplexans, 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 macroschizont to maintain host-cell transformation.

References and Notes
2. L. M. Forsyth et al., Comp. Pathol. 120, 39 (1999).
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 database with consecutive accession numbers between CR940346 and CR940353; they can be viewed at www.genei.ebi.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

www.sciencemag.org SCIENCE VOL 309 1 JULY 2005

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 ± SE. EST data were from cDNAs from three life-cycle stages in TA (macroschizont, 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 GPls occurred in the absence of signal peptides, it was 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/GPl proteins versus other merozoite proteins (P = 0.0166; 95% CI: 0.0024 to 0.2080), Mann-Whitney test; ‡, macroschizont/signal/NLS proteins versus other macroschizont proteins (P = 0.001; 95% CI: 0.0483 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.
Editor's Summary

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

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