DATABASE The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID bar015, doi:10.1093/database/bar015

Original article

TparvaDB: a database to support *Theileria parva* vaccine development

Paul Visendi[1], Wanjiku Ng'ang'a[2], Wallace Bulimo[3], Richard Bishop[4], James Ochanda[1] and Etienne P. de Villiers[4,*]

[1]Center for Biotechnology and Bioinformatics, University of Nairobi, Nairobi, [2]School of Computing and Informatics, University of Nairobi, Nairobi, [3]US Army Medical Research Unit – Kenya, Nairobi and [4]International Livestock Research Institute, Nairobi, Kenya

***Corresponding author. Tel:** +254204223000; Fax: +254204223001; Email: e.villiers@cgiar.org

Submitted 6 January 2011; Revised 23 March 2011; Accepted 10 April 2011

We describe the development of TparvaDB, a comprehensive resource to facilitate research towards development of an East Coast fever vaccine, by providing an integrated user-friendly database of all genome and related data currently available for *Theileria parva*. TparvaDB is based on the Generic Model Organism Database (GMOD) platform. It contains a complete reference genome sequence, Expressed Sequence Tags (ESTs), Massively Parallel Signature Sequencing (MPSS) expression tag data and related information from both public and private repositories. The Artemis annotation workbench provides online annotation functionality. TparvaDB represents a resource that will underpin and promote ongoing East Coast fever vaccine development and biological research.

Database URL: http://tparvadb.ilri.cgiar.org

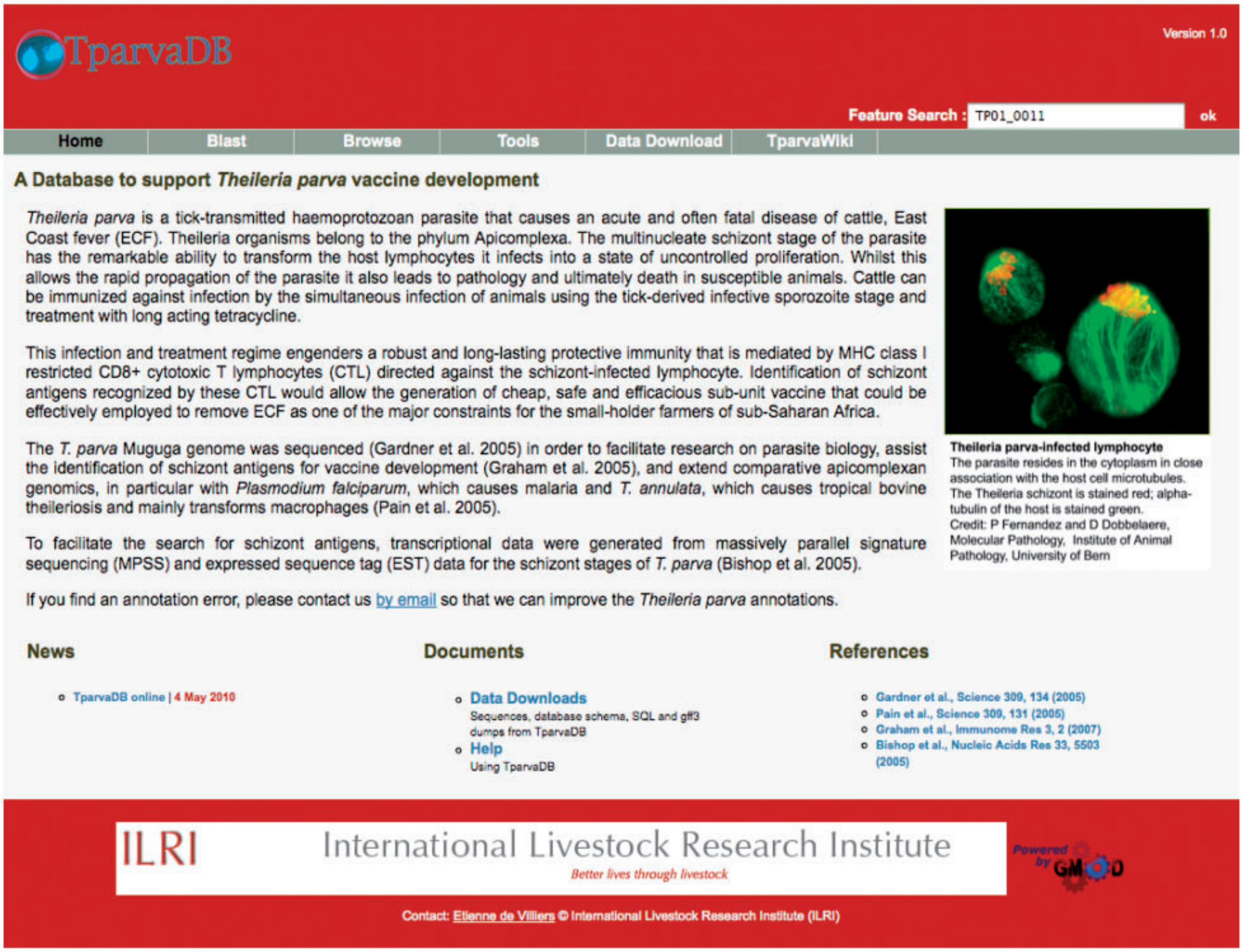
Introduction

Theileria parva is a tick-transmitted haemoprotozoan parasite that causes an acute and often fatal disease of cattle, East Coast fever (ECF) (1). ECF severely constrains the livelihoods of poor livestock-keepers in sub-Saharan Africa. Current control methods include use of acaricides to limit tick populations, drug-treatment of cattle exhibiting clinical symptoms and deployment of a live vaccine that involves infection with a potentially lethal dose of cryo-preserved sporozoites and simultaneous treatment with long-acting oxy-tetracycline (2). A subunit vaccine will provide a long-term solution to this socio-economically important constraint to livestock development in Eastern and Southern Africa.

The completion of the genome sequence of *T. parva* Muguga (3) represents an important milestone in research on the parasite biology, and has contributed to the identification of candidate schizont antigens for vaccine development targeting this stage of the parasite (4). It is also an important resource for apicomplexan comparative genomics, in particular with *Plasmodium falciparum*, which causes malaria in humans and *T. annulata* (5), the cause of tropical bovine Theileriosis, a related disease of cattle that has a wide range in North Africa and South and East Asia. To date the utilization of the *T. parva* genome and associated information by interested scientists has been limited by the lack of a user-friendly interface that provides access not only to genome data but the large set of expression data from MPSS (6) and EST data for the schizont stage and additional unpublished and published microarray expression data (7). In addition, the current system of access does not easily allow for updating of annotation and curation as new data becomes available (8).

TparvaDB architecture

To address the need to provide a comprehensive resource to facilitate research in the development of an ECF vaccine and comparative Apicomplexan genomics, TparvaDB was developed to provide a unified platform for all genome and related *T. parva* data (Figure 1). TparvaDB was built using components of the Generic Model Organism Database (GMOD) system (http://www.gmod.org).



A Database to support *Theileria parva* vaccine development

Theileria parva is a tick-transmitted haemoprotozoan parasite that causes an acute and often fatal disease of cattle, East Coast fever (ECF). Theileria organisms belong to the phylum Apicomplexa. The multinucleate schizont stage of the parasite has the remarkable ability to transform the host lymphocytes it infects into a state of uncontrolled proliferation. Whilst this allows the rapid propagation of the parasite it also leads to pathology and ultimately death in susceptible animals. Cattle can be immunized against infection by the simultaneous infection of animals using the tick-derived infective sporozoite stage and treatment with long acting tetracycline.

This infection and treatment regime engenders a robust and long-lasting protective immunity that is mediated by MHC class I restricted CD8+ cytotoxic T lymphocytes (CTL) directed against the schizont-infected lymphocyte. Identification of schizont antigens recognized by these CTL would allow the generation of cheap, safe and efficacious sub-unit vaccine that could be effectively employed to remove ECF as one of the major constraints for the small-holder farmers of sub-Saharan Africa.

The *T. parva* Muguga genome was sequenced (Gardner et al. 2005) in order to facilitate research on parasite biology, assist the identification of schizont antigens for vaccine development (Graham et al. 2005), and extend comparative apicomplexan genomics, in particular with *Plasmodium falciparum*, which causes malaria and *T. annulata*, which causes tropical bovine theileriosis and mainly transforms macrophages (Pain et al. 2005).

To facilitate the search for schizont antigens, transcriptional data were generated from massively parallel signature sequencing (MPSS) and expressed sequence tag (EST) data for the schizont stages of *T. parva* (Bishop et al. 2005).

If you find an annotation error, please contact us by email so that we can improve the *Theileria parva* annotations.

Theileria parva-infected lymphocyte
The parasite resides in the cytoplasm in close association with the host cell microtubules. The Theileria schizont is stained red; alpha-tubulin of the host is stained green.
Credit: P Fernandez and D Dobbelaere, Molecular Pathology, Institute of Animal Pathology, University of Bern

News

- TparvaDB online | 4 May 2010

Documents

- Data Downloads
  Sequences, database schema, SQL and gff3 dumps from TparvaDB
- Help
  Using TparvaDB

References

- Gardner et al., Science 309, 134 (2005)
- Pain et al., Science 309, 131 (2005)
- Graham et al., Immunome Res 3, 2 (2007)
- Bishop et al., Nucleic Acids Res 33, 5503 (2005)

ILRI
International Livestock Research Institute
Better lives through livestock
Powered by GMOD

Contact: Etienne de Villiers © International Livestock Research Institute (ILRI)

Figure 1. Screenshot of TparvaDB interface providing access to several online tools.

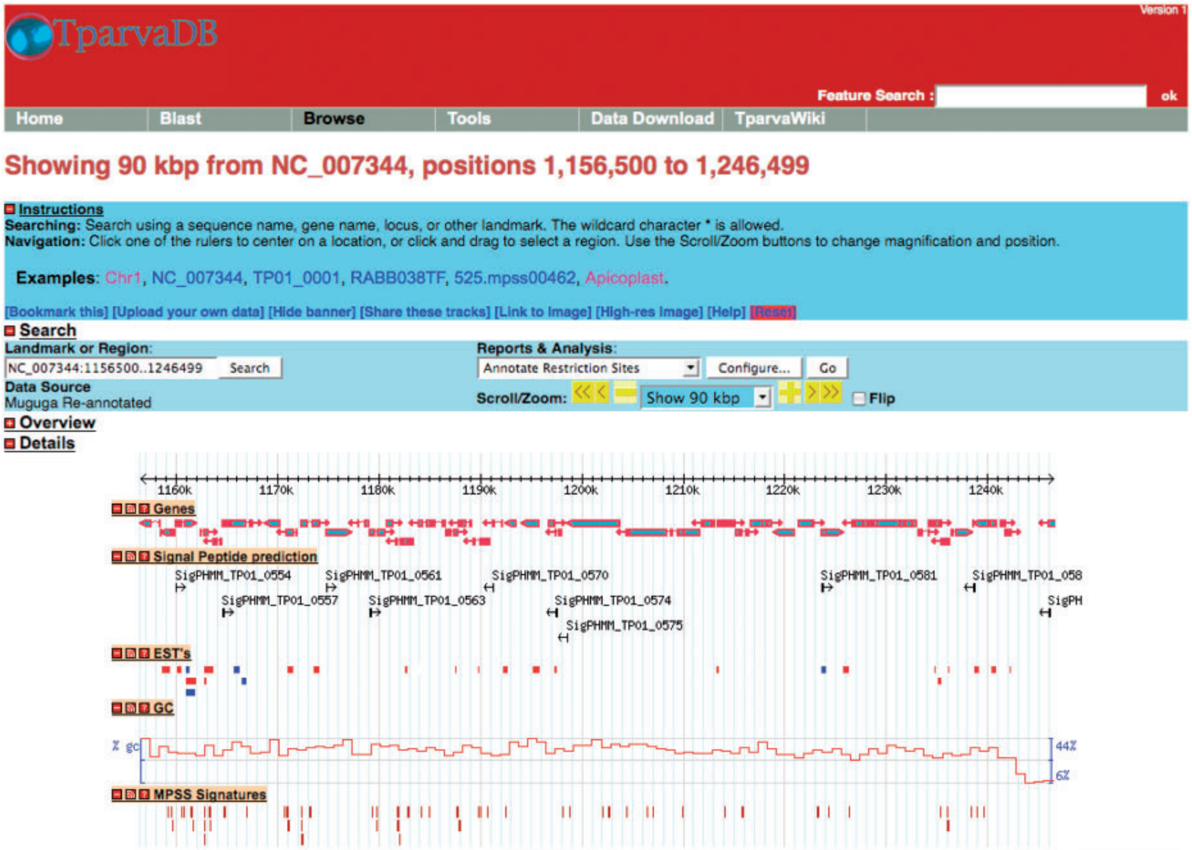
The core component of TparvaDB is Chado, an ontology driven relational database schema (9), using the PostgreSQL open source database management system (http://www.postgresql.org). Ontologies used in TparvaDB include Sequence (10), Gene (11) and Relationship (12) ontologies.

Theileria parva Muguga genome data in GenBank format was converted to GFF3 format and loaded into the TparvaDB Chado database using Perl scripts. EST, MPSS and SignalP data was similarly converted to GFF3 format and uploaded to the database. Several tools were included in TparvaDB to provide functionality. These include, GBrowse, NCBI Blast and an online annotation tool Artemis (13).

GBrowse, is a platform independent, extensible web-based graphical interface application for the visualization of genomic features stored in a database (14). In TparvaDB, GBrowse utilizes the Bio::DB::Das::Chado perl interface to connect directly to the TparvaDB Chado database, and the Bio::Graphics perl libraries to create images. GBrowse allows a simultaneous genome overview combined with the facility for a detailed view of specific regions of the genome. A user can query chromosomal regions of interest and visualise specific features such as annotated genes, ESTs and MPSS data mapped to a specific region (Figure 2). Online data analysis plug-ins provides detailed information on specific features that can be downloaded.

The NCBI Blast (15) tool allows homology searches against either nucleotide or protein sequences from *T. parva* (Muguga). In future, database releases data for other *T. parva* isolates that are currently being sequenced, and will be included as they are completed.

We implemented an online annotation functionality using Artemis, a DNA sequence viewer and annotation tool (13). Artemis supports Chado databases through the use of the iBATIS DataMapper API (http://ibatis.apache.org), allowing real-time connection to the underlying TparvaDB Chado database. Users with granted permissions can login to the Chado database remotely and make changes to *T. parva* gene models online through the Artemis interface (Figure 3).



Showing 90 kbp from NC_007344, positions 1,156,500 to 1,246,499

Instructions

Searching: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.

Navigation: Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.

Examples: Chr1, NC_007344, TP01_0001, RABB038TF, 525.mpss00462, Apicoplast.

[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [High-res Image] [Help] [Reset]

Search

Landmark or Region: NC_007344:1156500..1246499 Search

Data Source: Muguga Re-annotated

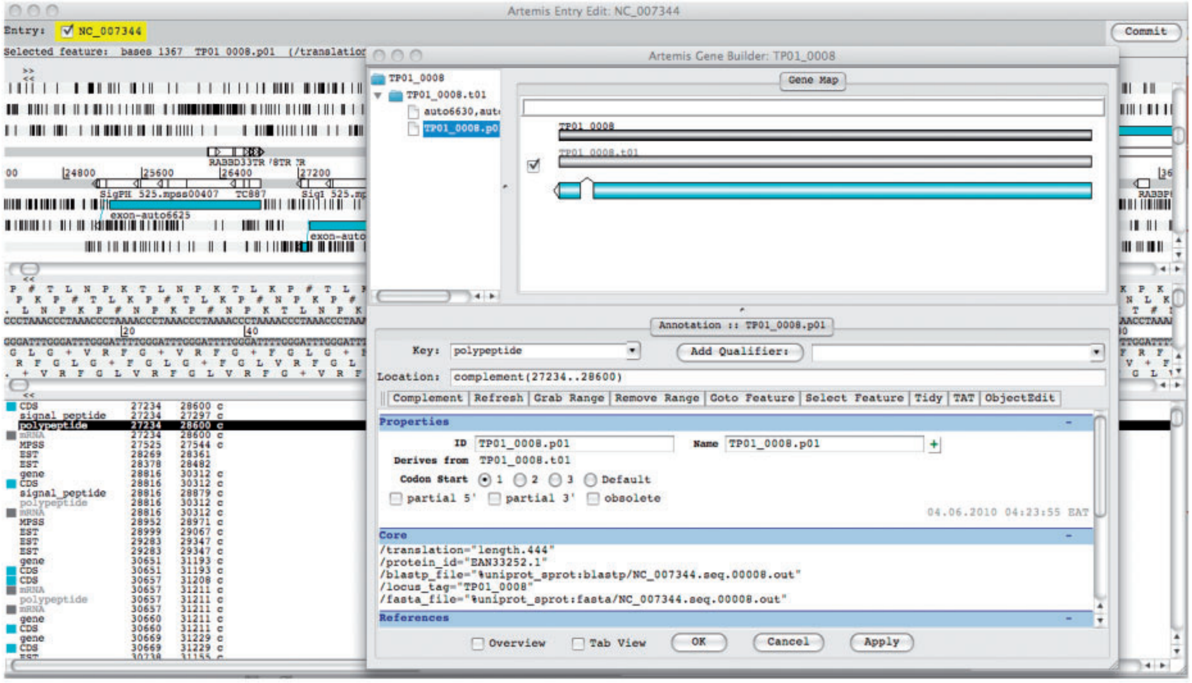
Reports & Analysis: Annotate Restriction Sites Configure... Go

Scroll/Zoom: Show 90 kbp □ Flip

Overview

Details

Figure 2. A representative example of the GBrowse interface. Tracks display information on annotated genes, predicted signal peptides, mapped ESTs, %GC content and MPSS signatures in a region of Chromosome 1.



```
P # T L N P K T L N P K T L K P # T L
 P K P # T L K P # T L K P # N P K P #
. L N P K P # N P K P # N P K T L N P K
CCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAA
                 20                  40
GGGATTTGGGATTTGGGATTTGGGATTTGGGATTTGGGATTTGGGATT
 G L G + V R F G + V R F G + V R F G L
  R F G L G + F G L G + F G L V R F G L
. + V R F G L V R F G L V R F G + V R F
```

```
CDS              27234  28600 c
signal_peptide   27234  27297 c
polypeptide      27234  28600 c
mRNA             27234  28600 c
MPSS             27525  27544 c
EST              28269  28361
EST              28378  28482
gene             28816  30312 c
CDS              28816  30312 c
signal_peptide   28816  28879 c
polypeptide      28816  30312 c
mRNA             28816  30312 c
MPSS             28952  28971 c
EST              28999  29067 c
EST              29283  29347 c
EST              29283  29347 c
gene             30651  31193 c
CDS              30651  31193 c
CDS              30657  31208 c
mRNA             30657  31211 c
polypeptide      30657  31211 c
mRNA             30657  31211 c
gene             30660  31211 c
CDS              30660  31211 c
gene             30669  31229 c
CDS              30669  31229 c
EST              30738  31155 c
```

Annotation :: TP01_0008.p01

Key: polypeptide Add Qualifier:

Location: complement(27234..28600)

Complement | Refresh | Grab Range | Remove Range | Goto Feature | Select Feature | Tidy | TAT | ObjectEdit

Properties

ID TP01_0008.p01 Name TP01_0008.p01 +

Derives from TP01_0008.t01

Codon Start ● 1 ○ 2 ○ 3 ○ Default

□ partial 5' □ partial 3' □ obsolete

04.06.2010 04:23:55 EAT

Core

```
/translation="length.444"
/protein_id="EAN33252.1"
/blastp_file="%uniprot_sprot:blastp/NC_007344.seq.00008.out"
/locus_tag="TP01_0008"
/fasta_file="%uniprot_sprot:fasta/NC_007344.seq.00008.out"
```

References

□ Overview □ Tab View OK Cancel Apply

Figure 3. Screenshot of Artemis annotation tool allowing direct access to *T. parva* genome data in the TparvaDB database for online annotation. Background image depicts Artemis interface and foreground image depict Gene Builder view.

Table 1. Annotation changes based on EST-based re-annotation

Chromosome	Gene model	Change	Comment
1	TP01_0115	Modified CDS	
1	TP01_0289	Modified CDS	Exon added
1	TP01_0869	Modified CDS	Exon added
1	TP01_1252	New CDS	
1	TP01_1253	New CDS	
1	TP01_1254	Modified CDS	Exon added
1	TP01_1255	New CDS	
2	TP02_0127	Modified CDS	Exon added
2	TP02_0140	Modified CDS	Exon added
2	TP02_0173	Modified CDS	Exon added
2	TP02_0181	Modified CDS	Exon added
2	TP02_0209	Modified CDS	Exon added
2	TP02_0236	Modified CDS	Exon added
2	TP02_0247	Modified CDS	Exon added
2	TP02_0942	Modified CDS	Exon added
2	TP02_0980	New CDS	
2	TP02_0981	New CDS	
2	TP02_0982	New CDS	
2	TP02_0983	New CDS	
2	TP02_0984	New CDS	
2	TP02_0986	New CDS	
3	TP03_0029	Modified CDS	Exon added
3	TP03_0041	Modified CDS	Exon added
3	TP03_0224	Modified CDS	Four exons added
3	TP03_0349	Modified CDS	Exon added
3	TP03_0413	Modified CDS	Exon added
3	TP03_0542	Modified CDS	Exon added
3	TP03_0574	Modified CDS	Exon added
3	TP03_0792	Modified CDS	Exon added
3	TP03_0862	Modified CDS	Exon added
3	TP03_0945	New CDS	
3	TP03_0946	New CDS	
4	TP04_0943	New CDS	
4	TP04_0566	Modified CDS	Exon added
4	TP04_0945	New CDS	
4	TP04_0843	Modified CDS	Exon added

Theileria parva re-annotation

As mentioned several additional *T. parva* isolates are currently being sequenced which will enable comparative genomics analysis to aid ECF vaccine research. One of the most effective methods to annotate a newly sequenced genome is to compare it with a well-annotated closely related genome using computational tools and databases. In order to improve the quality of the reference genome, we re-annotated T. *parva* Muguga genome using data that had become available since the original annotation, most importantly a large EST data set generated since the first genome was sequenced. In this exercise, we used TparvaDB to re-annotate the *T. parva* Muguga genome using the new data. *Theileria parva* schizont and sporozoite EST data were downloaded from NCBI dbEST database (16), mapped to the *T. parva* genome using both PASA (17) and BLAT (18), and results loaded into TparvaDB. The original *T. parva* genome data and the mapped EST data were visualized and gene models manually annotated using Artemis. In total, 13 new gene models were created and 23 gene models were corrected as shown in Table 1.

Future development of *T. parva* DB

To expedite selection of potential vaccine targets, and to determine the impact and potential limitations of extensive vaccine deployment, additional *T. parva* genomes are being sequenced. These new genomes can be quickly annotated using TparvaDB as they become available. We plan to incorporate a comparative genomics component to TparvaDB to facilitate identification of conserved and divergent regions between isolates. The comparative genomics component could either be implemented through GBrowse_syn (http://gmod.org/wiki/GBrowse_syn), a GBrowse-based synteny browser or by interfacing TparvaDB with other comparative database systems such as Sybil (http://sybil.sourceforge.net/). In addition, the ECF vaccine development project is generating immunological data on additional candidate antigens, bovine class I MHC sequences and other immunological data that can also be incorporated in TparvaDB. To manage this increasingly sophisticated information more efficiently, we plan to implement a query system based on BioMart, a query-oriented data management system that provides 'data mining'-like searches of complex descriptive data (19).

Funding

Multilateral core funding provided to the International Livestock Research Institute supported this work. This is ILRI publication number IL-201101. Funding for open access charge: International Livestock Research Institute.

Conflict of interest. None declared.

References

1. Norval,R.A.I., Perry,B.D. and Young,A.S. (1992) *The Epidemiology of Theileriosis in Africa*. Academic Press, London.

2. Radley,D.E. (1981) *Infection and Treatment Method of Immunization Against Theileriosis.* Martinus Nijhoff Publishers, The Hague.
3. Gardner,M.J., Bishop,R., Shah,T. *et al*. (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, **309**, 134–137.
4. Graham,S.P., Pelle,R., Honda,Y. *et al*. (2006) *Theileria parva* candidate vaccine antigens recognized by immune bovine cytotoxic T lymphocytes. *Proc. Natl Acad. Sci. USA*, **103**, 3286–3291.
5. Pain,A., Renauld,H., Berriman,M. *et al*. (2005) Genome of the host-cell transforming parasite Theileria annulata compared with *T. parva*. *Science*, **309**, 131–133.
6. Bishop,R., Shah,T., Pelle,R. *et al*. (2005) Analysis of the transcriptome of the protozoan *Theileria parva* using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. *Nucleic Acids Res.*, **33**, 5503–5511.
7. Schmuckli-Maurer,J., Casanova,C., Schmied,S. *et al*. (2009) Expression analysis of the *Theileria parva* subtelomere-encoded variable secreted protein gene family. *PLoS One*, **4**, e4839.
8. Shah,T., de Villiers,E., Nene,V. *et al*. (2006) Using the transcriptome to annotate the genome revisited: application of massively parallel signature sequencing (MPSS). *Gene*, **366**, 104–108.
9. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
10. Eilbeck,K., Lewis,S.E., Mungall,C.J. *et al*. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
11. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**, 25–29.
12. Smith,B., Ceusters,W., Klagges,B. *et al*. (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
13. Carver,T., Berriman,M., Tivey,A. *et al*. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.
14. Stein,L.D., Mungall,C., Shu,S. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Rodriguez-Tome,P. (1997) Searching the dbEST database. *Methods Mol. Biol.*, **69**, 269–283.
17. Haas,B.J., Delcher,A.L., Mount,SM. *et al*. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
18. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
19. Smedley,D., Haider,S., Ballester,B. *et al*. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.