Highly syntenic and yet divergent: a tale of two *Theilerias*

Willie Weir[1], Jack Sunter[2], Marie Chaussepied[3], Robert Skilton[2], Andrew Tait[1], Etienne P. de Villiers[2], Richard Bishop[2], Brian Shiels[1] and Gordon Langsley[3*]

[1]Institute of Comparative Medicine, Glasgow University Veterinary School, Bearsden Road, Glasgow, G61 1QH, UK.

[2]International Livestock Research Institute, PO Box 30709, Nairobi, Kenya.

[3]Institut Cochin, Inserm, U567, Cnrs, UMR 8104, Faculté de Médecine, Université Paris Descartes - Hôpital Cochin, 27, rue du Faubourg Saint-Jacques, 75014 Paris, France.

*Corresponding author: gordon.langsley@inserm.fr

Abstract

The published genomic sequences of the two major host-transforming *Theileria* species of cattle represent a rich resource of information, which has allowed novel bioinformatic and experimental studies into these important apicomplexan parasites. Since their publication in 2005, the genomes of *T. annulata* and *T. parva* have been utilised for a diverse range of applications, ranging from candidate antigen discovery to the identification of genetic markers for population analysis. This has led to advancements in the quest for a sub-unit vaccine, while providing a greater understanding of variation among parasite populations in the field. The unique ability of these *Theileria* species to induce host cell transformation is the subject of considerable scientific interest and the availability of full genomic sequences has also provided new insights into this area of research. This article reviews data underlying published comparative analyses, focussing on the general features of gene expression, the major *Tpr*/*Tar* multi-copy gene family and a re-examination of the predicted macroschizont secretome. Codon usage between the *Theileria* species is reviewed in detail, as this underpins ongoing comparative studies investigating selection at the intra- and inter-species level. The *TashAT*/*TpshAT* family of genes, conserved between *T. annulata* and *T. parva*, encodes products targeted to the host nucleus and has been implicated in contributing to the transformed bovine phenotype. Species-specific expansion and diversification at this critical locus is discussed, with reference to the availability, in the near future, of genomic datasets which are based on non-transforming *Theileria* species.

.

Introduction

Theileria annulata and *T. parva* are apicomplexan parasites that infect and transform bovine leukocytes, causing a widespread leukaemia-like disease of great economic importance, reviewed in Dobbelaere & Rottenberg, 2003. In the field, their tropism for specific host leukocytes differs, with *T. parva* infecting T cells, whereas *T. annulata* invades mostly cells of myeloid origin (monocytes and macrophages). It is possible that the difference in host cell predilection may underlie differences in the pathology of the two diseases, with T cell-based East Coast Fever (ECF) caused by *T. parva* being more pernicious than the myeloid-based tropical theileriosis caused by *T. annulata*. Over and above their veterinary clinical relevance, the ability of *T. annulata* and *T. parva* to reversibly induce leukocyte transformation, recently reviewed in Heussler et al., 2006, has also generated much interest due to potential mechanistic (epigenetic) similarities with some aspects of human cancer. In spite of the heavy economic cost caused by the two diseases in large parts of the Old World, the existing live vaccines are only delivered on a limited scale. The infection and treatment method of vaccination against ECF induces a persistent 'carrier' state (Oura et al., 2004), although the significance of this is unclear (McKeever, 2007). The unique ability of these two apicomplexa to induce host cell transformation combined with the economically important disease syndromes that they induce provided the rationale for determination of the complete genome sequence of both *T. parva* and *T. annulata* and subsequent comparative analysis (Pain et al., 2005; Gardner et al., 2005). Mining this data set for parasite genes with the potential to modulate host cell phenotypes has been reviewed recently by Shiels et al., 2006.

The published genomes of *T. annulata* and *T. parva* have already been used in comparative studies and shed new light on the biology of apicomplexan parasites in general. For example, it has been shown that across the apicomplexa, species-specific genes posses stronger bias in codon usage compared to other genes in the genome with a large number of genus or species-specific genes encoding putative surface antigens (Kuo & Kissinger, 2008). In *Theileria*, surface antigen genes

are conserved at the genus level and distributed across chromosomes, contrasting with *Plasmodium* where antigen encoding genes are located in the sub-telomeres and are largely species-specific (Kuo & Kissinger, 2008). With the advent of the published genome sequence of *T. parva*, it has been possible to screen the genome *in silico* for genes encoding a secretory signal peptide and this has facilitated a targeted approach to functional screening of T cell antigen candidates. Of the 986 predicted genes on *T. parva* chromosome I, a subset of 55 was predicted to encode secreted antigens. 36 of these genes were cloned and together with a series of random schizont cDNA clones, they were used in an immuno-screening approach to identify MHC I-presented antigens (Graham et al., 2007). Comparative genomics using *Theileria* has also facilitated antigen discovery in other apicomplexan species. *Babesia* and *Theileria* show extensive conservation of synteny and using positional analysis, the putative orthologue of the major *Theileria* sporozoite surface antigen SPAG-1/p67 has been identified in *B. bovis* (Brayton et al., 2007).

In addition to studies focusing on those genes encoding antigens and putatively secreted proteins, comparative analyses of non-coding regions of the genome have yielded interesting results. *T. annulata* and *T. parva* were shown to share 99.7 % of intron positions in conserved regions of the genome, with both species being considerably more intron-rich than *P. falciparum* (Roy & Penny, 2006). A further comparative study revealed that the common ancestor of *Theileria* and *Plasmodium* probably contained more introns than extant *Theileria* species, indicating intron-loss is outstripping intron-gain (Roy & Penny, 2007). More importantly, a genome-wide analysis of intergenic regions of each *T. annulata* and *T. parva* identified a number of conserved motifs (Guo & Silva, 2008). This included two putative transcription factor binding sites, thus providing candidates for experimental investigation. The availability of complete genomic sequences has also allowed the development of a new generation of genetic markers for population studies. Using a preliminary assembly of the genome, a panel of 11 micro-satellite and 49 mini-satellite polymorphic markers was identified in *T. parva* (Oura et al., 2003). These fast evolving loci were found to be distributed across all four chromosomes, predominantly in non-coding regions and

following experimental validation all were shown to be specific to *T. parva*. A panel of ten markers was later identified in the genome of *T. annulata* (Weir et al., 2007) and both marker sets have since been applied to population genetic studies (Oura et al., 2005; Weir et al., 2007).

Here, we extend and review the comparative analysis of these two highly syntenic genomes with particular attention to general features of gene expression, the major *Tpr* and *Tar* multi-copy gene families and a re-examination of the predicted macroschizont secretome including genes encoding putative peptidases.

Codon usage in *T. annulata* and *T. parva* compared to *Plasmodium*

Non-synonymous substitutions are defined as single nucleotide changes in DNA sequence, which encode a variant amino acid. In contrast, synonymous substitutions do not result in amino acid change. The rate of non-synonymous (d_N) to synonymous substitutions (d_S), referred to as $d_N d_S$ is a useful index for quantifying the influence of purifying and diversifying selection. We have performed a genome-wide analysis of codon usage and bias within and between the two *Theilerias* and *Plasmodium*, as this will underpin new comparative genomics and validate previous stage-specific $d_N d_S$ analysis (Pain et al., 2005). The software package CodonW (http://codonw.sourceforge.net/) was used to calculate indices of Relative Synonymous Codon Usage (RSCU), which measures the ratio of the observed frequency of a codon relative to that expected if codon usage is uniform, with values tending towards one indicating an absence of bias.

The coding sequence of *T. annulata* comprises 2,030,707 codons and RSCU values for this dataset were calculated (Table 1.i). Certain codons are encountered more frequently than others and the codon with the highest RSCU value is AGA, encoding arginine (Arg) with a value of 2.95. This contrasts with the lowest values for CGG (0.17) and CGC (0.29), two of the other five codons encoding this residue. For each other amino acid encoded by more than one codon, the value for the most frequent codon ranges between 1.18 for CAU (histidine) to 1.95 for UCA (serine). To assess whether codon usage differs between species, RSCU was calculated for the 1,902,549 codons in the *T. parva* genome (Table 1.ii). Similar to *T. annulata*, the amino acid with the most

divergent RSCU was arginine with values ranging from 2.67 (AGA) to 0.24 (CGG) and RSCU values ranging between 1.03 and 1.80 were observed for the preferred codons CAU and UCA with the trend of RSCU towards a particular subset of codons was identical. RSCU values in *T. annulata* were slightly more polarised than in *T. parva* and this is demonstrated in Figure 1, where RSCU of synonymous codons is correlated between *T. annulata* and *T. parva* and between *T. annulata* and *P. falciparum*. Linear regression for the *T. annulata* / *T. parva* plot has a gradient near, but less than, unity (0.83), underpinning the observation that preferred codons in *T. annulata* are also preferred in *T. parva*, but not quite to the same extent. This contrasts with the relationship of codon usage of *T. annulata* in comparison to *P. falciparum*. The subset of preferred codons in *P. falciparum* is similar, but their usage is greater than in *T. annulata* (linear regression gradient of 1.42) and this is probably related to the greater AT-richness of the coding sequences of *Plasmodium* compared with *Theileria*.

RSCU was calculated over all genes with expressed sequence tag (EST) data in the *T. annulata* genome and codon preference was found to be almost identical across the three stages (Supplementary Figure 1). Correspondence analysis was performed on the coding sequences of *T. annulata*, allowing differentiation of genes based on codon usage. The most highly and least biased genes were identified as the 5 % of genes at either extreme of the principal axis and where RSCU for a particular codon was greater in the most biased compared to the least biased subset, a codon was identified as putatively optimal (Supplementary Table 1). Using this methodology, putatively optimal codons were found to be broadly similar between species, across the different bovine stages of *T. annulata* and within the subset of genes that comprise the *T. annulata* secretome (Table 2).

Importantly, as there is no evidence of differential codon usage between each species and between life-cycle stages, meaningful comparisons can be made using these datasets for inter-species and intra-species $d_N d_S$ studies. This observation supports an earlier study, which determined that elevated $d_N d_S$ values are associated with predicted merozoite surface antigens (Pain et al., 2005).

Codon usage by *T. annulata* is clearly not random, a common observation across the majority of eukaryotic species. It has been demonstrated in other organisms that the major factor explaining bias in codon usage between genes is the expression level of the encoded protein, with highly expressed genes using a limited subset of codons (Sharp & Matassi, 1994). This hypothesis will be partially addressed in the near future through micro-array analysis of parasite gene expression at the mRNA level in *T. annulata*, although further proteomic studies will be required to test whether the effect applies at the level of translation.

A relatively high proportion of species-specific genes encode secreted products

On comparing species-specific genes to those genes common between *T. annulata* and *T. parva*, a higher proportion of species-specific genes were found to encode signal peptides and this observation is reflected across each of the three life-cycle stages (Table 3). For genes expressed in the merozoite and/or piroplasm, a much higher proportion of species-specific genes encode trans-membrane domains, many of which are *Tar/Tpr* genes without defined orthologues. In this study, species-specific genes are identified where a one-to-one orthologous relationship cannot be established by reciprocal BLAST analysis. In some cases, species-specific genes occur at the same locus in each species, show sequence similarity and are members of related gene families. It may be postulated that species-specific genes have largely arisen since *T. annulata* and *T. parva* diverged, whereas conserved genes were present in the common ancestor and these are more likely to be represented in other *Theileria* species. The finding that many species-specific genes are predicted to encode proteins that are secreted into the macroschizont-infected cell implies that through speciation, novel parasite genes have been selected, the products of which encode a signal peptide and perform their function within the host cell compartment. Of these 42 genes, four genes encode multiple trans-membrane domains and likely correspond to integral membrane proteins, including one member of the *Tar* family which is constitutively expressed. Twelve of the remaining 38 genes represent members of the sub-telomerically-encoded variable secreted protein (SVSP) family (Pain et al., 2005) and with the exception of *TashAT1* and *TashAT3*, all the other

genes are annotated as encoding hypothetical proteins. Interestingly, using the predictNLS algorithm (Cokol et al., 2000), only the two TashAT genes and one SVSP gene are found to encode a nuclear localisation signal. TashAT family proteins have been implicated in controlling the host cell phenotype (Swan et al., 1999; Swan et al., 2001; Shiels et al., 2004) and the comparative genomics of this family is addressed later in this review. In contrast, the SVSP family has not been experimentally characterised and its function is currently unknown, although its genomic location together with its variability within each species is compatible with a role in immune evasion (Barry et al., 2003). The postulation that species-specific genes encoding secreted/host-interacting products have been selected is supported by the relatively high level of $d_N d_S$ computed for genes of this class (Pain et al., 2005), while novel merozoite and piroplasm-expressed products, and products of merozoite orthologous gene pairs that display high $d_N d_S$, are biased toward a trans-membrane location. However, both polypeptides secreted by the macroschizont and proteins on the surface of the merozoite/piroplasm may be more likely to be exposed to the protective immune response than other classes of proteins expressed by these stages. The resulting selection pressure in combination with gene duplication may therefore have generated gene sequences that are too divergent to be matched to orthologues by reciprocal BLASTing. This may account for the observation that the large families of genes encoding proteins with signal peptides in the two genomes often contain members that lack an orthologue and have therefore been defined as species-specific in this study.

Shared and species-specific genes show different RNA expression profiles

In order to compare gene expression profiles of the two species, it was necessary to utilise transcriptional data generated by different techniques. In *T. annulata*, transcriptional profiling of the different life-cycle stages is based on ESTs that were derived from sequencing around 500bp of individually cloned stage-specific cDNAs made from merozoites, piroplasms and macroschizonts, with the largest collection of tags coming from macroschizonts (Pain et al., 2005). In *T. parva*, expression profiling was done exclusively on macroschizonts and generated by MPSS, massively

parallel signature sequencing, a PCR-based technique that gives short (20bp) sequence tags of very high coverage (Bishop et al., 2005). Initially, the two data sets were compared by analysing all shared genes with both EST and MPSS sense data (n = 2216) and a correlation co-efficient of 0.234 was calculated (*p = 0.000)*. In part, the low co-efficient value reflects the vastly different sizes of the two data sets. Transcriptional profiles of shared versus species-specific and constitutive versus stage-specific genes throughout the life-cycle were also compared (Table 4). The remarkably high counts of species-specific piroplasm ESTs are largely due to the highly expressed *Tar / Tpr* genes without definite orthologues. For both species, a higher percentage of the common gene-set is expressed in the macroschizont compared to the species-specific gene-sets. Moreover, *T. parva*-specific genes transcribed in this stage are expressed at a lower level than shared genes and, since MPSS data has been $\log_2$-transformed, the statistically significant difference of 5.58 versus 6.44 ($p < 0.001$, Mann-Whitney test) is suggestive of an almost a two-fold difference in transcriptional activity.

The MPSS approach generates both sense and anti-sense data for a given gene. Anti-sense transcripts have been described for 12 % of *P. falciparum* genes and notably, sense and anti-sense tag counts from single loci across the transcriptome were inversely related, leading to the suggestion that anti-sense transcripts may play some negative regulatory role (Gunasekera et al., 2004). Similarly, 14 % of *T. parva* genes have anti-sense transcripts (Bishop et al., 2005). We therefore analysed the 398 genes where both signals were present, comparing sense and anti-sense levels, however no correlation was identified. We then compared gene-sets with anti-sense transcripts to those without and found the MPSS signal to be higher in the group with corresponding anti-sense data. This may suggest that in *Theileria,* in contrast to *Plasmodia,* these more highly expressed genes or networks of genes require anti-sense regulation or perhaps that anti-sense transcripts act to regulate gene expression at the level of translation. Interestingly, evidence for translational control of the gene encoding the major merozoite/piroplasm surface antigen Tams1 has been reported in the macroschizont stage (Swan et al., 2001).

To test if highly expressed genes are evolving at a higher rate than other genes in the *Theileria* genomes, $d_N d_S$ ratios were compared to MPSS and EST scores for each gene. We found that $d_N d_S$ values are inversely correlated with macroschizont expression data, whether it was MPSS (correlation co-efficient: -0.198, $p < 0.001$, n = 2216), or ESTs (correlation co-efficient: -0.0402, $p = 0.022$, n = 3250). For MPSS, the relationship is stronger and more significant, probably because of better distribution and larger dynamic range of the dataset compared with the EST collection. Therefore, with increasing $d_N d_S$, transcriptional activity decreases and this is evidence that rapidly evolving genes tend not to be as highly expressed as the more slowly evolving genes. Consequently, the $d_N d_S$ data are consistent with the previous observation that genes conserved between species have higher levels of transcription compared to specifies-specific genes. The simplest explanation for this result is that genes shared by both species are more likely to confer a conserved function that requires abundant levels of protein, although again a proteomic study would be necessary to investigate this hypothesis.

The *Tpr* and *Tar* multi-copy gene families of *T. parva* and *T. annulata*

Despite the high level of genomic synteny between the two species, a striking difference between the *T. parva* and *T. annulata* genomes is the arrangement and expression of a rapidly evolving multi-copy gene family, designated '*Tpr*' in *T. parva,* and '*Tar*' in *T. annulata* (Pain et al., 2005; Gardner et al., 2005). The organisation of *Tpr* is summarised in Figure 2, Panel 1 and that of *Tar* in Figure 2, Panel 2. The conserved feature that defines the *Tpr*/*Tar* genes is an approximately 260 amino acid domain encoded at the 3′ end of all the open reading frames (ORFs). This core C-terminal domain is frequently, but not invariably, associated with additional regions of potential protein coding sequence that exhibit different levels of repetition in the genome. The *Tpr*/*Tar* genes are normally conserved as ORFs, with coding potential. The *Tar* genes are dispersed throughout the *T. annulata* genome. In contrast, the *T. parva* genome has a tandem array of *Tpr* genes containing a minimum of 28 ORFs (Gardner et al., 2005) , although due to its complexity the entire Tpr locus has not been assembled. *T. parva* also contains twelve ORFs dispersed over each of the

four chromosomes and combining the tandemly arrayed and dispersed copies, there are 40 defined *Tpr* genes. One hypothesis is that both the expansion in number of the dispersed *Tar* copies and generation of the *Tpr* array has occurred post-speciation. Alternatively, it is conceivable that ancestral dispersed copies in the gene family of *T. parva* have been lost following speciation. Previous sequence analysis identified three putative repeated protein segments Tpr3, Tpr2 and Tpr1 arranged in the order 3, 2, 1, from the N- to C-terminus within a region of the Tpr locus (Baylis et al., 1991). These repeated segments contained a high concentration of trans-membrane domains (TMDs). The TMDs within the Tpr1 domain were the most conserved sections of the predicted protein when different strains were analysed, whereas divergence was observed in the regions between the TMD (Bishop et al., 1997). Only five *Tpr* genes contain all three domains (Tpr3, Tpr2, and Tpr1) and 14 do not contain a Tpr2 domain; when all domains are present their orientation is always 3, 2, 1, or 2 followed by 1 when only Tpr2 and Tpr1 are present together. In contrast, in *Tar* genes, predicted polypeptides possessing all three domains are much more frequent, 69 of the 84 dispersed *Tar* genes exhibit a 3, 2, 1 domain structure (organisation of predicted *Tar* genes is summarised in Figure 2, Panel 2 and Table 5). The presence of multiple TMDs within the Tpr and Tar predicted proteins suggest a membrane location. In addition, analysis of expression patterns has indicated that mRNAs representing dispersed *Tpr* genes are associated with the transcriptome of the macroschizont (Bishop et al., 2005) while genes of the tandem array are more likely to be transcribed by the intra-erythrocytic piroplasm stage transmitted to the tick (Bishop et al., 1997). Unusually, more than 50 % of ORFs within a tandemly arrayed section of the Tpr locus lack an in-frame ATG codon within the first 50 amino acids of the ORF, suggesting an unusual expression mechanism. Clearer understanding of the observed diversity of these gene families, both between and within *Theileria* species, will require detailed study of the proteins they encode and a greater insight into their putative function.

Signal peptide analysis of the *T. annulata* proteome

Theileria parasites alter their host leukocyte's signal transduction programme and this is believed to underpin host cell transformation. One predicted mechanism is secretion of molecules into the host cell cytosol that modify leukocyte signal transduction pathways and for this reason much attention has been paid to identifying putative host-transforming factors secreted macroschizont stage of the parasite (Pain et al., 2005), recently reviewed in Shiels et al., 2006.

Given the importance of correctly estimating the *Theileria* secretome, we decided to re-examine the predicted *T. annulata* proteome using two different algorithms - a neural network (NN) and a Hidden Markov Model (HMM), and we have summarised the results using both algorithms (Supplementary Table 2). Using the HMM, there are 448 proteins predicted to have a secretory signal and 198 with a signal anchor i.e. with a recognised signal peptide, but without a cleavage site; the latter have not been included in summary sheet. An additional 180 secreted proteins are identified with the NN algorithm, which are included in the summary sheet. Generally, the results of the two methods agree fairly well. We re-examined *T. annulata* genes annotated as coding for kinases (79) and phosphatases (28) and one that was annotated as potentially coding for both, giving 106 putative proteins in total. Two phosphatases and one kinase have a secretory signal predicted by SignalP3.0 and were predicted in our previous SignalP2.0 analysis (Pain et al., 2005). *Toxoplasma gondii* has been shown to secrete kinases into its host cell and alter the gene expression programme (Taylor et al., 2006; Saeij et al., 2007): the single *Theileria* kinase (TA09960 - putative cell-cycle-related serine/threonine protein kinase, CDK homologue) and two phosphatases (TA07270 - putative proton translocating inorganic pyrophosphatase and TA04960 - putative acid phosphatase) merit further analysis. All three genes have orthologues in *T. parva,* denoted as TP04_0791, TP04_0216 and TP03_0512. The kinase and acid phosphatase are well conserved between species with amino acid identities of 90 % and 92 % respectively and $d_N d_S$ indicates they are both under purifying selection. In contrast, the pyrophosphate shows lower identity at the protein level and evidence of positive selection, with an above average $d_N d_S$ value of 0.1594. In addition to kinases and phosphatases, peptidases that are active in the host

compartment of the macroschizont-infected cell could contribute to the transformed phenotype (Pain et al., 2005; Shiels et al., 2006). Further analysis of peptidases with signal peptides *in silico*, confirmed their potential to be transported across membranes. However, unequivocal identification of peptidases predicted to be secreted into the host cell and specifically expressed by the macroschizont was not achieved. A strong prediction that an expanded family of membrane-bound cysteine protease genes play a role in determining the transformed phenotype is also not possible as they display elevated expression in infected cells undergoing merozoite production, a process associated with inhibition of leukocyte proliferation, or are expressed by the intra-erythrocytic piroplasm stage. In contrast, re-analysis of the dataset highlighted a membrane peptidase, TA18300/Tp03_0804, that is expressed specifically by the macroschizont stage at high levels. This protein may have an important indirect role in establishment of the transformed phenotype: it is predicted to cleave signal peptides from proteins, thus allowing translocation of macroschizont secretome proteins into the host compartment.

The *TashAT*/*TpshAT* gene families

The TashAT family of genes in *T. annulata* encodes secreted proteins that translocate to the host nucleus and bind DNA. This gene family is clearly replicated in *T. parva* with evidence of eight direct orthologous gene pairs. However, the two gene families also display significant species-specific diversification (Shiels et al., 2006). These findings have been confirmed by further analysis showing that synteny across four orthologous gene pairs located at either end of the cluster maintains gene order and that this extends across the intergenic regions (Figure 3). In contrast, genes internal to the cluster do not show the same order across species. Phylogenetic analysis of the gene families also highlights that diversification has occurred at this locus within each species. Sequences of the internal genes cluster separately forming two clades representing species-specific genes, whereas genes located at the termini show most similarity to their orthologues compared to paralogous sequences within their respective families (Figure 4). It would appear, therefore, that the TashAT family was condensed in the common ancestor and that it

has undergone significant expansion and diversification as the species diverged. Thus, genes flanking the cluster are more likely to perform a function conserved across the species while internal genes may have evolved during species diversification, including adaptation to a preferred host cell type. A predicted function for TashAT family proteins is that they act as DNA binding cofactors, which direct the expression of genes targeted by bovine transcription factors (such as AP1 and NF-kB) constitutively activated following infection of the leukocyte by both species (Oura et al., 2006).

Conclusions

- **Codon usage in *T. annulata* and *T. parva* compared to *Plasmodium*.** Codon usage is almost identical in the genomes of *T. annulata* and *T. parva* showing clear bias towards particular codons, an observation common to the majority of eukaryotes. The subset of preferred codons in *Theileria* is similar to *Plasmodia*, but their usage is less polarised, probably reflecting the greater AT-richness of *P. falciparum* coding sequences compared to *Theileria*.

- **A relatively high proportion of species-specific genes encode secreted products.** The finding that many species-specific proteins were predicted to be secreted into the cytosol of the macroschizont-infected leukocyte implies that during speciation of *T. annulata* and *T.* parva these genes have diversified as each species adapted to a particular biological niche. In addition, this location may promote exposure to a protective immune response resulting in pressure that allows selection of divergent forms. In both cases the pressure to diversify may have resulted in the generation of gene families undergoing rapid expansion via gene duplication, genetic recombination and diversification of amino acid sequence.

- **Shared and species-specific genes show different RNA expression profiles.** The distribution of expression data among genes common to *T. annulata* and *T. parva* versus species-specific genes showed that a lower percentage of species-specific genes are expressed in macroschizonts. Transcription profiles of conserved versus species-specific and constitutive versus stage-specific genes throughout the life-cycle also showed higher expression is associated

with more conserved genes in all parasite stages. Moreover, $d_N d_S$ data suggest that conserved genes have higher levels of transcription compared to specifies-specific implying that genes shared by both species are more likely to confer a conserved function that requires abundant levels of protein.

- **Anti-sense transcription.** Anti-sense transcripts have been described for 12 % of *P. falciparum* genes and similarly, 14 % of *Theileria* genes have anti-sense transcripts. The group of *Theileria* genes with anti-sense transcripts has higher sense MPSS levels, and this could suggest that in *Theileria,* in contrast to *Plasmodia,* these more highly expressed genes / networks might require anti-sense regulation.

- **The *Tpr* and *Tar* multi-copy gene families of *T. parva* and *T. annulata*.** The *Tpr* locus has the features of a system that has evolved for the generation of diversity and there are at least 28 *Tpr* genes organised in a tandem array located centrally on chromosome III. The tandem array is absent from *T. annulata*, but *Tar* genes dispersed throughout the genome of *T. annulata* are considerably more numerous than the twelve dispersed *Tpr* copies in *T. parva*. It seems likely that the common ancestor must have contained an ancestral form of *Tpr* and *Tar* and both creation of the tandem array in *T. parva* and expansion and divergence of single copy loci have occurred. The function of the species-specific diversity in *Tar*/*Tpr* gene families is not clear and requires further study.

- **Signal peptide analysis of the *T. annulata* proteome.** *Theileria* parasites alter their host leukocyte's signal transduction programme and induce cellular transformation. One predicted mechanism is via secretion of parasite-encoded kinases or phosphatases into the host cell cytosol. Surprisingly, only two phosphatases and one kinase were predicted to have a secretory signal making them candidates worthy of future study at the functional level. Interestingly, *Theileria* encode a single membrane peptidase that is specifically and strongly expressed at the macroschizont stage. This protease, which is predicted to cleave signal peptides from proteins,

could play an important indirect role in establishing the transformed phenotype by allowing translocation of macroschizont secretome proteins into the host compartment.

- **The *TashAT/TpshHN* gene families.** These two families have expanded and diverged, and this may have occurred during speciation of *T. annulata* and *T. parva*. Therefore, it seems likely that the family encoded by the common ancestor species comprised a significantly smaller number of genes. In regard to this postulation it will be interesting to analyse the genome of a non-transforming *Theileria* species to determine if an orthologous family exists and whether it is significantly condensed. The most likely function for proteins encoded by the *TashAT/TpshHN* families is that they operate to tailor the profile of host genes that are expressed by the infected leukocyte, possibly as cofactors to host transcription factors that are constitutively activated by the parasite. If so divergence of the *TashAT/TpshHN* families may explain the observed differences in genes expressed by *T. annulata* and *T. parva* infected cells (Sager et al., 1998), even though evidence suggests that both species activate the same bovine transcription factors (Dobbelaere & Kuenzi, 2004).

Table 1. Relative synonymous codon usage

(i) *T. annulata* genome (2,030,707 codons)

AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU
Phe	UUU	62973	1.25	Ser	UCU	35836	1.23	Tyr	UAU	54680	1.24	Cys	UGU	23079	1.38
	UUC	37619	0.75		UCC	19087	0.65		UAC	33568	0.76		UGC	10327	0.62
Leu	UUA	55727	1.64		UCA	56825	1.95	TER	UAA	2733	2.17	TER	UGA	529	0.42
	UUG	41339	1.21		UCG	12950	0.44		UAG	516	0.41	Trp	UGG	16493	1.00
	CUU	35287	1.04	Pro	CCU	23693	1.26	His	CAU	24128	1.18	Arg	CGU	8570	0.66
	CUC	20376	0.60		CCC	10175	0.54		CAC	16823	0.82		CGC	3716	0.29
	CUA	28827	0.85		CCA	34332	1.82	Gln	CAA	41273	1.32		CGA	5387	0.42
	CUG	22904	0.67		CCG	7134	0.38		CAG	21119	0.68		CGG	2208	0.17
Ile	AUU	62914	1.31	Thr	ACU	45710	1.50	Asn	AAU	100053	1.29	Ser	AGU	35633	1.22
	AUC	25223	0.52		ACC	20663	0.68		AAC	54634	0.71		AGC	14661	0.5
	AUA	56163	1.17		ACA	44528	1.46	Lys	AAA	98626	1.21	Arg	AGA	38088	2.95
Met	AUG	41639	1.00		ACG	10973	0.36		AAG	64700	0.79		AGG	19623	1.52
Val	GUU	47299	1.58	Ala	GCU	22312	1.32	Asp	GAU	81245	1.39	Gly	GGU	29885	1.31
	GUC	16262	0.54		GCC	13331	0.79		GAC	35606	0.61		GGC	11263	0.5
	GUA	33846	1.13		GCA	26683	1.58	Glu	GAA	89349	1.36		GGA	40025	1.76
	GUG	22380	0.75		GCG	5323	0.31		GAG	42047	0.64		GGG	9787	0.43

(ii) *T. parva* genome (1,902,549 codons)

AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU
Phe	UUU	58772	1.25	Ser	UCU	33500	1.21	Tyr	UAU	45036	1.11	Cys	UGU	21439	1.36
	UUC	35149	0.75		UCC	20755	0.75		UAC	36425	0.89		UGC	10104	0.64
Leu	UUA	48298	1.5		UCA	49699	1.80	TER	UAA	3001	2.21	TER	UGA	588	0.43
	UUG	39790	1.24		UCG	13058	0.47		UAG	489	0.36	Trp	UGG	15090	1.00
	CUU	31783	0.99	Pro	CCU	23928	1.28	His	CAU	21115	1.03	Arg	CGU	8599	0.66
	CUC	23119	0.72		CCC	12543	0.67		CAC	19883	0.97		CGC	4204	0.32
	CUA	25540	0.79		CCA	29378	1.57	Gln	CAA	37123	1.22		CGA	5047	0.39
	CUG	24709	0.77		CCG	9170	0.49		CAG	23757	0.78		CGG	3202	0.24
Ile	AUU	54831	1.32	Thr	ACU	41151	1.47	Asn	AAU	80406	1.17	Ser	AGU	34997	1.27
	AUC	25851	0.62		ACC	21784	0.78		AAC	57368	0.83		AGC	13546	0.49
	AUA	43906	1.06		ACA	37144	1.33	Lys	AAA	85578	1.15	Arg	AGA	35013	2.67
Met	AUG	38557	1.00		ACG	11979	0.43		AAG	63081	0.85		AGG	22482	1.72
Val	GUU	45566	1.54	Ala	GCU	21096	1.3	Asp	GAU	71221	1.28	Gly	GGU	25486	1.20
	GUC	16943	0.57		GCC	14708	0.90		GAC	40257	0.72		GGC	14417	0.68
	GUA	29265	0.99		GCA	22186	1.36	Glu	GAA	73387	1.21		GGA	33305	1.57
	GUG	26543	0.90		GCG	7040	0.43		GAG	47522	0.79		GGG	11640	0.55

AA = encoded amino acid, n = number of codons, RSCU = relative synonymous codon usage

Legend Table 1.

Relative synonymous codon usage (RSCU) values were calculated across three datasets – **(i)** the *T. annulata* genome, **(ii)** the *T. parva* genome and **(iv)** the *P. falciparum* genome. RSCU measures the ratio of the observed frequency of a codon relative to that expected if codon usage is uniform, i.e. values tending towards 1.00 indicate an absence of bias. Amino acids encoded by a single codon necessarily have a RSCU value of 1.00. The number of times each codon is encountered in the dataset (n) is recorded, including stop codons (TER). Standard three-letter codes are used to indicate amino acids.

Table 2. Comparison of putative optimal codons of *T. annulata* and *T. parva*

AA	codon	*T. annulata*	*T. parva*	Macro only	Mero only	Piro only	Secreted EST	AA	codon	*T. annulata*	*T. parva*	Macro only	Mero only	Piro only	Secreted EST
Phe	UUU	√	√	√	√	√	√	**Ser**	UCU	√	√	√	√	√	√
	UUC								UCC				*	√	
Leu	UUA	√	√	√	√	√	√		UCA		√				
	UUG								UCG						
	CUU		√		√	√		**Pro**	CCU	√	√	√	√	√	√
	CUC								CCC				*	*	
	CUA		√				*		CCA	√	√	√			*
	CUG								CCG						
Ile	AUU	√	*	*	√	√	†	**Thr**	ACU	√	√	√	√	√	√
	AUC								ACC					*	
	AUA	√	√	√	*		√		ACA		√				†
Met	AUG								ACG						
Val	GUU	√	√	√	√	√	*	**Ala**	GCU	√	√	√	√	√	√
	GUC					*			GCC						
	GUA	√	√	√	√		√		GCA		√	*			√
	GUG								GCG						
Tyr	UAU	√	√	√	√	√	√	**Cys**	UGU	√	√	√	√	√	√
	UAC								UGC						
TER	UAA							**TER**	UGA						
	UAG							**Trp**	UGG						
His	CAU	√	√	√	√	√	√	**Arg**	CGU	√		√	√	√	†
	CAC								CGC						
Gln	CAA	√	√	√	√	√	√		CGA		√			*	√
	CAG								CGG				*	*	
Asn	AAU	√	√	√	√	√	√	**Ser**	AGU	√	√	√	*	√	√
	AAC								AGC						
Lys	AAA	√	√	√	√	√	√	**Arg**	AGA	√	√	√	*		†
	AAG								AGG						
Asp	GAU	√	√	√	√	√	√	**Gly**	GGU	√	√	√	√	√	*
	GAC								GGC						
Glu	GAA	√	√	√	√	√	√		GGA		√				√
	GAG								GGG						

AA = encoded amino acid
√ $p < 0.01$, † $0.01 < p < 0.05$, * not statistically significant

Legend Table 2.

Correspondence analysis of codon usage was performed on the coding sequences comprising six datasets – (i) the entire *T. annulata* genome, (ii) the entire *T. parva* genome, (iii) macroschizont-specific *T. annulata* genes, (iv) merozoite-specific *T. annulata* genes, (v) piroplasm-specific *T. annulata* genes and (vi) *T. annulata* genes with a signal sequence and EST expression data. Subsets of the most highly and least biased genes were identified as the 5 % of genes at either extreme of the principal axis generated in each of the six analyses. Where the RSCU for a particular codon was greater in the most biased subset (High RSCU) compared to the least biased subset (Low RSCU), a codon was identified as putatively optimal.

Table 3. Features of shared genes and *T. annulata*-specific genes

Shared genes	Signal peptide	TMD	GPI	TA-specific genes	Signal peptide	TMD	GPI
Overall (3250)	443 (14%) *	635 (20%) *	60 (2%)	Overall (517)	124 (24%) *	200 (39%) *	13 (3%)
Macroschizont (1197)	167 (14%) *	257 (8%)	19 (0%)	Macroschizont (140)	42 (30%) *	49 (9%)	2 (0%)
Merozoite (667)	90 (13%) †	133 (4%) *	10 (0%)	Merozoite (122)	30 (25%) †	66 (13%) *	1 (0%)
Piroplasm (661)	75 (11%) *	114 (4%) *	10 (0%)	Piroplasm (109)	27 (25%) *	68 (13%) *	2 (0%)

* Chi-square, $p < 0.001$
† Chi-square, $p = 0.003$

Table 4. Variation in expression among different classes of gene

Number of genes expressed		Expression value	Number of genes expressed		Expression value	*p value* (difference in expression values *)
Shared vs species specific						
***T. annulata* shared genes (3250)**		**ESTs (median)**	***T. annulata* specific (517)**		**ESTs (median)**	
Macro	1197 (37%)	2	Macro	140 (27%)	3	0.151
Mero	667 (21%)	2	Mero	122 (24%)	3	<0.001
Piro	661 (20%)	2	Piro	109 (21%)	17	<0.001
***T. parva* shared genes (3250)**		**MPSS (mean)**	***T. parva* specific (753)**		**MPSS (mean)**	
Macro (S)	2216 (68%)	6.443	Macro (S)	302 (40%)	5.585	<0.001
Macro (AS)	464 (14%)	3.585	Macro (AS)	81 (11%)	3.907	0.346
Non-spliced vs spliced						
***T. annulata* non-spliced genes (1097)**		**ESTs (median)**	***T. annulata* spliced genes (2670)**		**ESTs (median)**	
Macro	367 (33%)	3	Macro	970 (36%)	2	<0.001
Mero	276 (25%)	3	Mero	513 (19%)	2	<0.001
Piro	275 (25%)	4	Piro	495 (18%)	2	<0.001
***T. parva* non-spliced genes (1025)**		**MPSS (mean)**	***T. parva* spliced genes (2979)**		**MPSS (mean)**	
Macro (S)	589 (57%)	6.087	Macro (S)	1923 (64%)	6.443	0.002
Macro (AS)	111 (11%)	3.585	Macro (AS)	429 (14%)	3.585	-
Constitutive vs stage-specific						
***T. annulata* constitutive**		**ESTs (median)**	***T. annulata* stage-specific**		**ESTs (median)**	
Macro	164	3.5	Macro	1173	2	<0.001
Mero	164	3	Mero	625	2	<0.001
Piro	164	4	Piro	606	2	<0.001

* Mann-Whitney Test

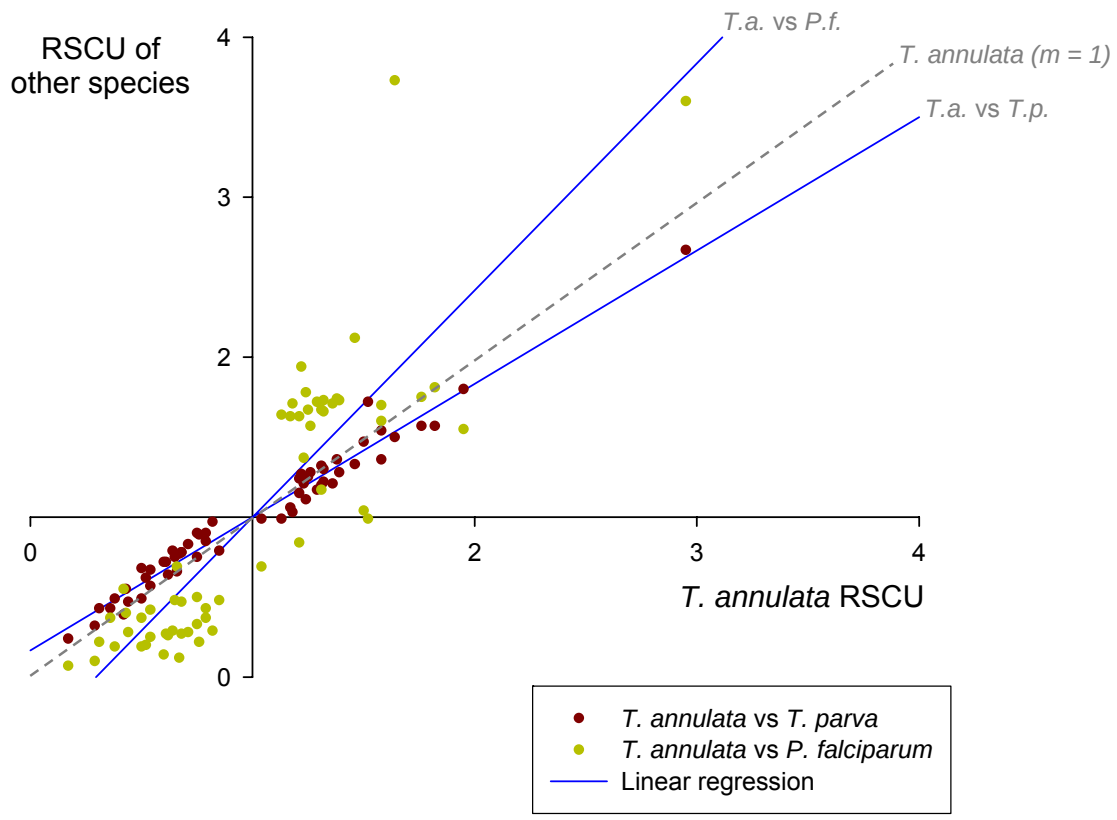
Table 5A. ORF lengths and levels of amino acid identity of five categories of *Tpr* ORF defined by presence of domain type (1-3) and arrangement

				Identity range (%)		
Type	**Number**	**Array**	**Length (aa)**	**Tpr3**	**Tpr2**	**Tpr1**
A	1	Y	796	n/a	n/a	n/a
B	4	N	512-1249	58-72	45-60	48-60
C	12	Y	606-738	not present	75-97	69-98
D	6	N	451-515	not present	30-35	33-37
E	14	Y	406-548	not present	not present	87-99

Table 5B. ORF lengths and levels of amino acid identity of the 69 *Tar* ORFs containing the Tar1, 2 and 3 domains

		Identity range (%)		
Number	**Length (aa)**	**Tar3**	**Tar2**	**Tar1**
69	432-1006	26-83	29-67	23-68

Figure 1. Correlation of relative synonymous codon usage of *T. annulata* vs *T. parva* and *T. annulata* vs *P. falciparum*

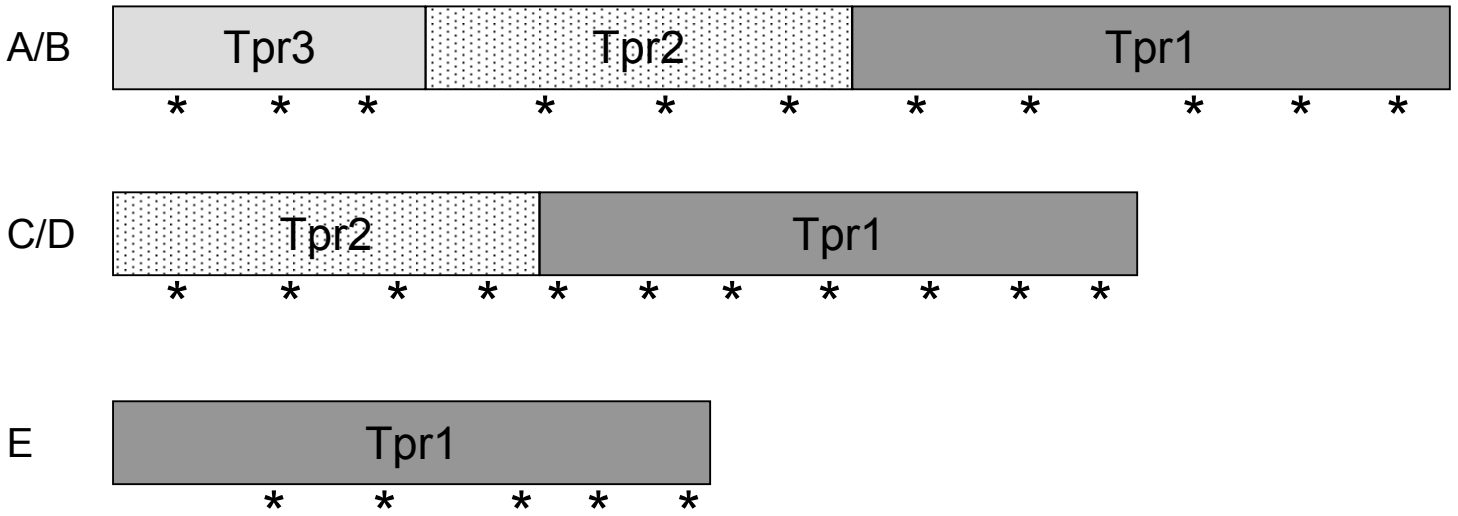


Legend Figure 1.

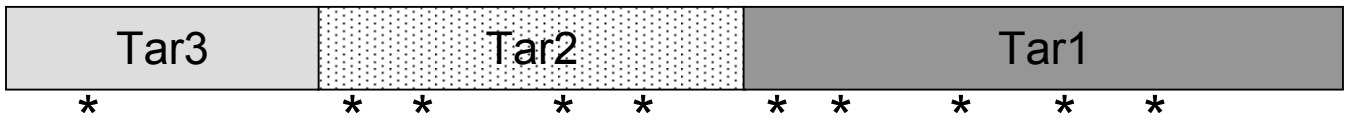
The relative synonymous codon usage (RSCU) of synonymous codons was correlated between *T. annulata* and *T. parva* (corresponding to the data contained in Table 1) and between *T. annulata* and *P. falciparum*. RSCU values of *T. annulata* are plotted on the x-axis while those from *T. parva* (**red**) and *P. falciparum* (**green**) are on the y-axis. Linear regression lines for each of the comparisons are marked in **blue**, with a dotted **grey** line representing a perfect match, i.e. *T. annulata* vs *T. annulata.*

Figure 2. Schematic representation of the organisation of repeated domains located in the *Tpr* and *Tar* genes of *Theileria*

Panel 1. *Tpr* genes



Panel 2. *Tar* genes

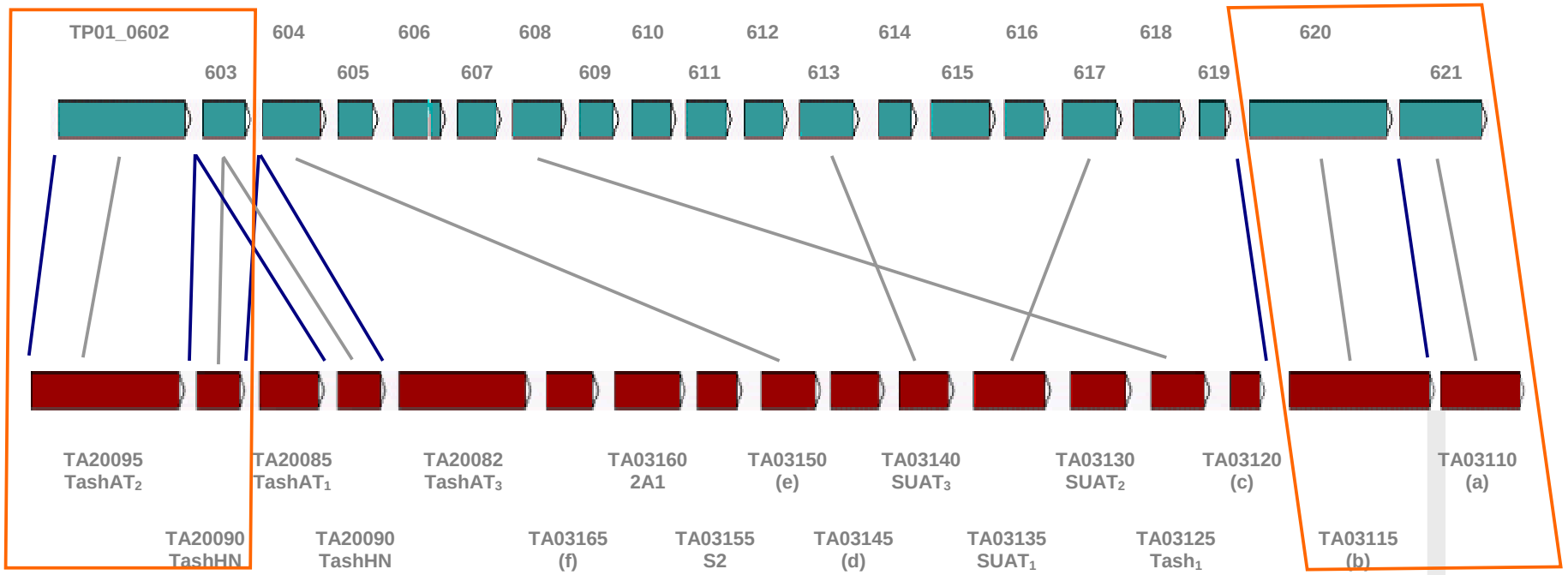


Legend Figure 2

Panel 1, Graphical illustration of categories of *Tpr* ORFs within the *T. parva* genome sequence according to arrangement of conserved protein domains; Panel 2, Graphical illustration of categories of *Tar* ORFs within the *T. annulata* genome according to the arrangement of conserved protein domains. Within The *Tpr* family A, C and E type genes are found on the array. B and D type genes are dispersed within the *T. parva* genome. The stars indicate the location of trans-membrane domains.

Figure 3. Synteny between *T. annulata* and *T. parva* at the TashAT locus

T. parva



T. annulata

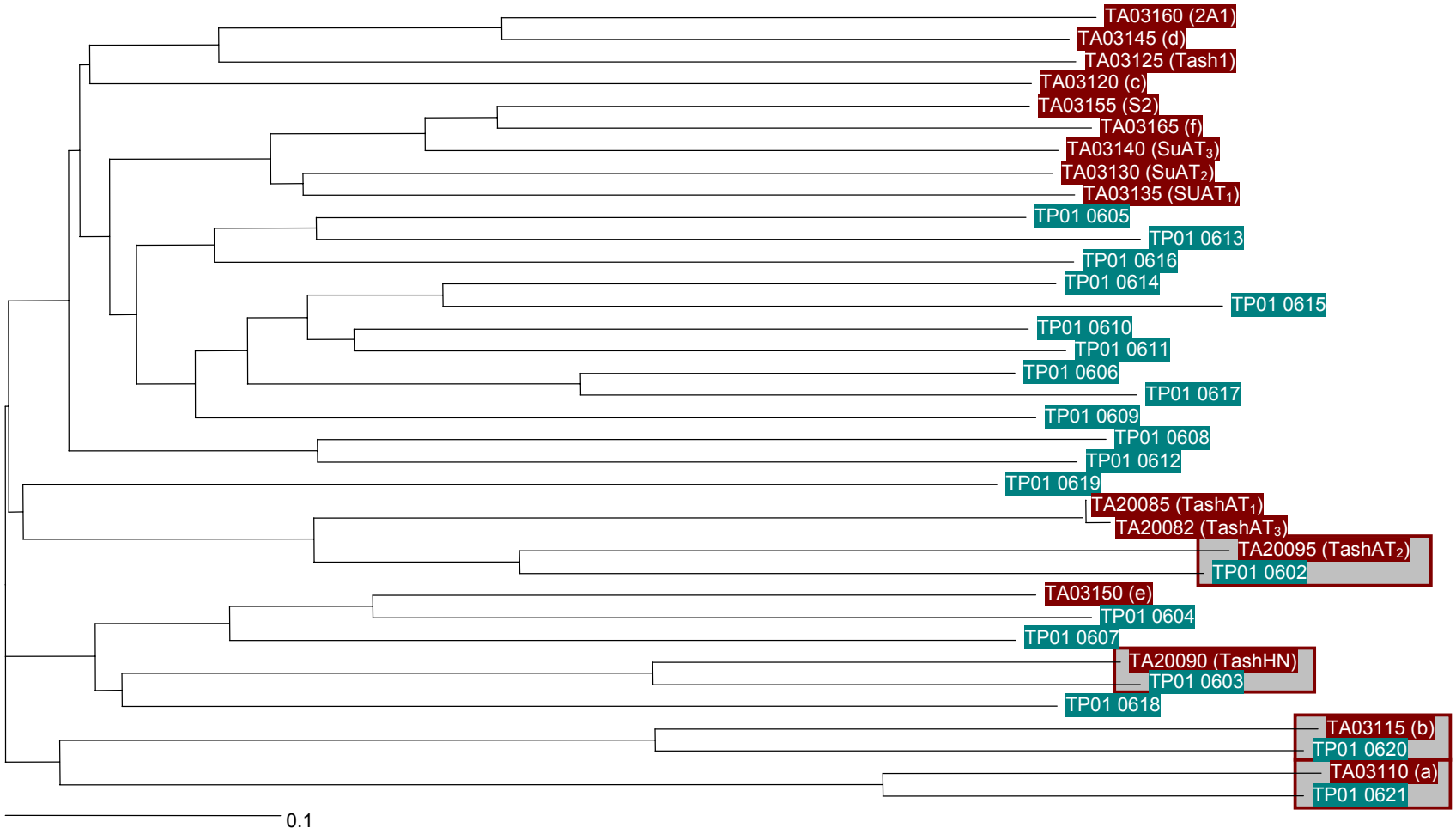
Intergenic region between TA03115(b) and TA03110(a) in *T. annulata* aligned with orthologous region in *T. parva* (74% identity)

```
TP01_0620   ACAGACAATGAATAAAACAACCATAATAATGACAATTAACTATAAATCGATATAAATTAATAATGAAATAAAACTAAATCAAGGCCAAAATATAGATAAATTAACGGGTAAT
TA03115(b)  TCAGATAATGAAGAGAA---------TAATAACAATTAAATATAAATTGAACCTAATTAATT-TAAAACAAGACT---TTAAAACAGTAATATC--TAAATTAATGTAAAAT
             **** ****** * **         **** ******** ******* **    *******  * *** ** ***   * **  *   *****   ******** *   ***

TAAGTGTGTAAATGTGGTTAGTTGTGAATTATAACAATTCGATGGATTAAGTCTGACTTCTTGGGGATTCTAACATATTTACAACAACACAAGCAGGTGATAATTAGAATGTATTAGTATAAT
TAAAAATG-AAATGAGGTTGGTTGTGTAATTTAAAAATTTGATGAATAGAGGTTAATTTCCTCAAGATTCCAACCAATTTACAACAACACATTTAAATGATAAGAGGAAAATACTAGCATAAT
***   ** ***** **** ****** * * *** **** **** **  **  * * *** *   ***** ***  ***************   *  ******   ***  ** *** *****

GAAGGATTAAGTATAAAAAATGTGTAAAAAAGATCCCCATATGCATTAGTAATAAAATTCGCATGTTTAGGTGTTATAATATGTTTTAATGTAATATATGTTTACACAT  TP01_0621
TAAGGATTAAGT-TAAAAAATGTGTAAAAACGATCCCCATATGCATTAACGATAAAAATCGCATGTTGAAGCGTTATAATATGTTTTAACTTAATATATGTTTACACAT  TA03110 (a)
*********** ***************** *****************   ****** ********* * * *****************  ******************
```

Legend Figure 3.

The TashAT locus in *T. annulata* is illustrated along with the orthologous locus in *T. parva*. Grey lines represent orthologous genes and blue lines represent orthologous intergenic regions. The two pairs of genes flanking each cluster have direct orthologues in the same position in the other species, together with conserved intergenic regions and are highlighted in orange.

Figure 4. Tree representing the TashAT family of *T. annulata* and orthologues in *T. parva*



Legend Figure 4.

The amino acid sequences of the sixteen genes in the TashAT family of T. annulata along with the twenty genes in the orthologous family in T. parva were aligned, clustered and used to create a tree. T. annulata genes are highlighted in red while T. parva genes are highlighted in blue. The orthologous genes corresponding to each end of the TashAT locus (Figure A) are highlighted in a grey box. The presence of two discrete clusters of internal genes suggests that these genes are more closely related to each other within a species and have evolved independently in each species. This is supported by the fact that most internal genes do not have direct orthologues (Figure 5).

Supplementary Table 1. Putative optimal codons of *T. annulata*

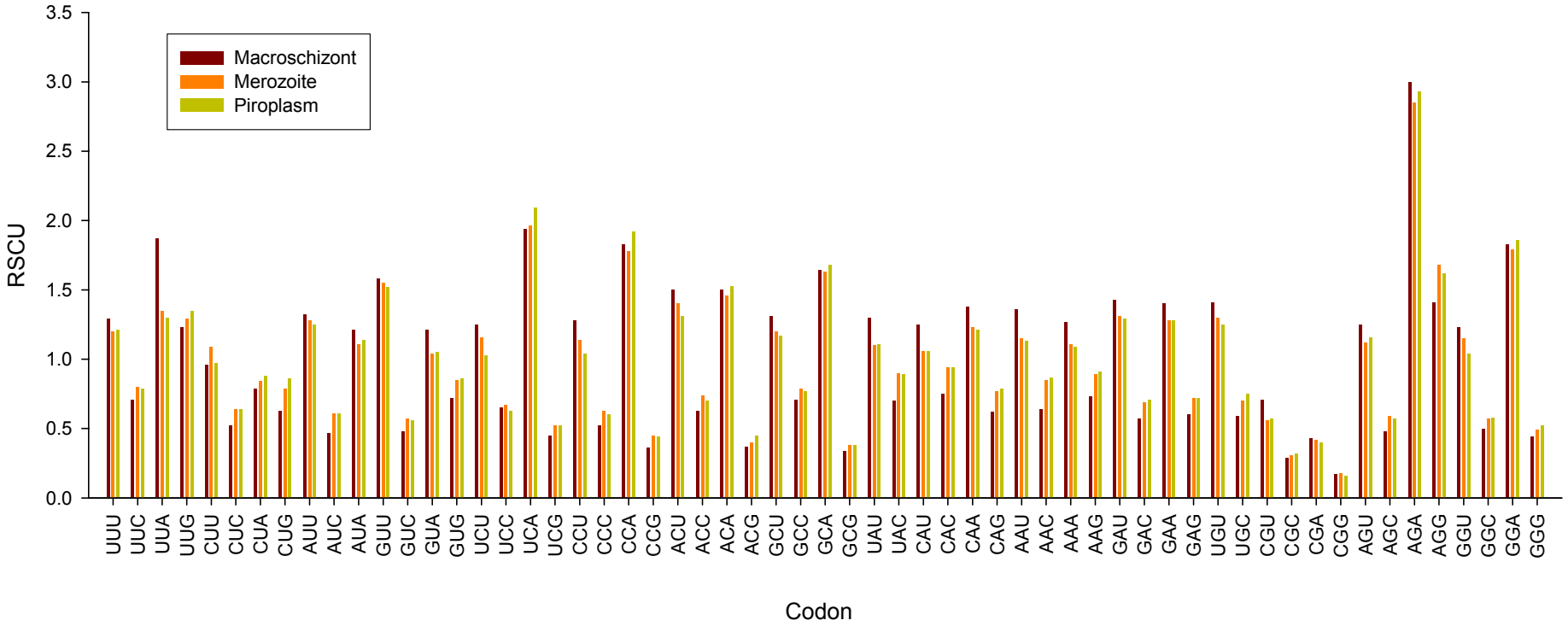
AA	codon	High RSCU	n	Low RSCU	n	AA	codon	High RSCU	n	Low RSCU	n
Phe						*Ser*					
	UUU *	1.69	(3754)	0.88	(1937)		UCU *	1.57	(1755)	0.90	(1137)
	UUC	0.31	(683)	1.12	(2450)		UCC	0.47	(532)	0.85	(1076)
Leu	UUA *	4.18	(6168)	0.72	(1106)		UCA	1.68	(1889)	1.95	(2461)
	UUG	0.69	(1024)	1.20	(1842)		UCG	0.14	(155)	0.67	(845)
	CUU	0.66	(976)	1.05	(1599)	Pro	CCU *	1.40	(933)	0.77	(660)
	CUC	0.11	(167)	1.09	(1675)		CCC	0.30	(201)	0.74	(632)
	CUA	0.30	(443)	0.83	(1272)		CCA *	2.18	(1455)	1.90	(1633)
	CUG	0.05	(74)	1.10	(1686)		CCG	0.12	(77)	0.59	(506)
Ile	AUU *	1.33	(4651)	1.13	(1802)	Thr	ACU *	2.08	(3790)	1.00	(1228)
	AUC	0.12	(415)	0.96	(1539)		ACC	0.44	(808)	0.94	(1155)
	AUA *	1.55	(5443)	0.91	(1456)		ACA	1.20	(2188)	1.52	(1868)
Met	AUG	1.00	(1827)	1.00	(2166)		ACG	0.27	(496)	0.55	(678)
Val	GUU *	1.66	(1491)	1.32	(2056)	Ala	GCU *	2.20	(927)	0.82	(858)
	GUC	0.13	(118)	0.84	(1308)		GCC	0.37	(158)	0.98	(1029)
	GUA *	1.79	(1607)	0.85	(1316)		GCA	1.36	(575)	1.79	(1883)
	GUG	0.41	(367)	0.99	(1537)		GCG	0.06	(26)	0.41	(430)
Tyr	UAU *	1.90	(4009)	0.73	(1316)	Cys	UGU *	1.93	(1384)	0.91	(627)
	UAC	0.10	(219)	1.27	(2286)		UGC	0.07	(52)	1.09	(749)
TER	UAA	2.39	(150)	2.27	(143)	TER	UGA	0.13	(8)	0.40	(25)
	UAG	0.48	(30)	0.33	(21)	Trp	UGG	1.00	(554)	1.00	(813)
His	CAU *	1.80	(1266)	0.70	(682)	Arg	CGU *	1.47	(570)	0.61	(414)
	CAC	0.20	(139)	1.30	(1259)		CGC	0.03	(11)	0.66	(444)
Gln	CAA *	1.86	(2127)	0.99	(1331)		CGA	0.31	(120)	0.44	(296)
	CAG	0.14	(159)	1.01	(1348)		CGG	0.11	(42)	0.14	(92)
Asn	AAU *	1.92	(10985)	0.74	(1997)	Ser	AGU *	1.94	(2175)	0.77	(968)
	AAC	0.08	(470)	1.26	(3366)		AGC	0.20	(221)	0.85	(1076)
Lys	AAA *	1.63	(6422)	0.83	(2711)	Arg	AGA *	3.56	(1385)	2.49	(1677)
	AAG	0.37	(1458)	1.17	(3785)		AGG	0.53	(206)	1.66	(1122)
Asp	GAU *	1.87	(3984)	0.93	(2342)	Gly	GGU *	1.87	(2104)	0.79	(829)
	GAC	0.13	(266)	1.07	(2710)		GGC	0.07	(78)	0.86	(903)
Glu	GAA *	1.62	(4858)	1.10	(3073)		GGA	1.83	(2062)	1.90	(2002)
	GAG	0.38	(1126)	0.90	(2522)		GGG	0.22	(252)	0.46	(490)

AA = encoded amino acid, n = number of codons, RSCU = relative synonymous codon usage
* $p < 0.01$

Legend Supplementary Table 1.

Correspondence analysis of codon usage was performed on the entire coding sequence of *T. annulata*. Subsets of the most highly and least biased genes were identified as the 5 % of genes at either extreme of the principal axis generated. Where the RSCU for a particular codon was greater in the most biased subset (High RSCU) compared to the least biased subset (Low RSCU), a codon was identified as putatively optimal.

Supplementary Figure 1. Relative synonymous codon usage across stage-specifically expressed genes for non-synonymous codons



Legend Supplementary Figure 1

The RSCU for non-synonymous codons was calculated for all genes in the *T. annulata* genome with EST information. The preference of particular codons is almost identical across the macroschizont (n = 736), merozoite (n = 279) and piroplasm (n = 168) stages of the life-cycle.

References

Barry, J.D., Ginger, M.L., Burton, P., McCulloch, R., 2003. Why are parasite contingency genes often associated with telomeres? Int. J. Parasitol. 33, 29-45.

Baylis, H.A., Sohal, S.K., Carrington, M., Bishop, R.P., Allsopp, B.A., 1991. An unusual repetitive gene family in *Theileria parva* which is stage-specifically transcribed. Mol. Biochem. Parasitol. 49, 133-142.

Bishop, R., Musoke, A., Morzaria, S., Sohanpal, B., Gobright, E., 1997. Concerted evolution at a multicopy locus in the protozoan parasite *Theileria parva*: extreme divergence of potential protein-coding sequences. Mol. Cell Biol. 17, 1666-1673.

Bishop, R., Shah, T., Pelle, R., Hoyle, D., Pearson, T., Haines, L., Brass, A., Hulme, H., Graham, S.P., Taracha, E.L., Kanga, S., Lu, C., Hass, B., Wortman, J., White, O., Gardner, M.J., Nene, V., de Villiers, E.P., 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.

Brayton, K.A., Lau, A.O., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, S.L., Brown, W.C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H.A., Haas, B.J., Howell, J.M., Khouri, H., Koo, H., Mann, D.J., Norimine, J., Paulsen, I.T., Radune, D., Ren, Q., Smith, R.K., Jr., Suarez, C.E., White, O., Wortman, J.R., Knowles, D.P., Jr., McElwain, T.F., Nene, V.M., 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS. Pathog. 3, 1401-1413.

Cokol, M., Nair, R., Rost, B., 2000. Finding nuclear localization signals. EMBO Rep. 1, 411-415.

Dobbelaere, D.A., Kuenzi, P., 2004. The strategies of the *Theileria* parasite: a new twist in host-pathogen interactions. Curr. Opin. Immunol. 16, 524-530.

Dobbelaere, D.A., Rottenberg, S., 2003. *Theileria*-induced leukocyte transformation. Curr. Opin. Microbiol. 6, 377-382.

Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A., Berriman, M., Wilson, R.J., Sato, S., Ralph, S.A., Mann, D.J., Xiong, Z., Shallom, S.J., Weidman, J., Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A.R., Wasawo, D., Crabtree, J., Wortman, J.R., Haas, B., Angiuoli, S.V., Creasy, T.H., Lu, C., Suh, B., Silva, J.C., Utterback, T.R., Feldblyum, T.V., Pertea, M., Allen, J., Nierman, W.C., Taracha, E.L., Salzberg, S.L., White, O.R., Fitzhugh, H.A., Morzaria, S., Venter, J.C., Fraser, C.M., Nene, V., 2005. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science 309, 134-137.

Graham, S.P., Honda, Y., Pelle, R., Mwangi, D.M., Glew, E.J., de Villiers, E.P., Shah, T., Bishop, R., van der, B.P., Nene, V., Taracha, E.L., 2007. A novel strategy for the identification of antigens that are recognised by bovine MHC class I restricted cytotoxic T cells in a protozoan infection using reverse vaccinology. Immunome. Res. 3, 2.

Gunasekera, A.M., Patankar, S., Schug, J., Eisen, G., Kissinger, J., Roos, D., Wirth, D.F., 2004. Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. Mol. Biochem. Parasitol. 136, 35-42.

Guo, X., Silva, J.C., 2008. Properties of non-coding DNA and identification of putative cis-regulatory elements in *Theileria parva*. BMC. Genomics 9, 582.

Heussler, V., Sturm, A., Langsley, G., 2006. Regulation of host cell survival by intracellular *Plasmodium* and *Theileria* parasites. Parasitology 132 Suppl, S49-S60.

Kuo, C.H., Kissinger, J.C., 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. BMC. Evol. Biol. 8, 108.

McKeever, D.J., 2007. Live immunisation against *Theileria parva*: containing or spreading the disease? Trends Parasitol. 23, 565-568.

Oura, C.A., Asiimwe, B.B., Weir, W., Lubega, G.W., Tait, A., 2005. Population genetic analysis and sub-structuring of *Theileria parva* in Uganda. Mol. Biochem. Parasitol. 140, 229-239.

Oura, C.A., Bishop, R., Wampande, E.M., Lubega, G.W., Tait, A., 2004. The persistence of component *Theileria parva* stocks in cattle immunized with the 'Muguga cocktail' live vaccine against East Coast fever in Uganda. Parasitology 129, 27-42.

Oura, C.A., McKellar, S., Swan, D.G., Okan, E., Shiels, B.R., 2006. Infection of bovine cells by the protozoan parasite *Theileria annulata* modulates expression of the ISGylation system. Cell Microbiol. 8, 276-288.

Oura, C.A., Odongo, D.O., Lubega, G.W., Spooner, P.R., Tait, A., Bishop, R.P., 2003. A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*. Int. J. Parasitol. 33, 1641-1653.

Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C.A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C., Cochet, M., Coulson, R.M., Cronin, A., de Villiers, E.P., Fraser, A., Fosker, N., Gardner, M., Goble, A., Griffiths-Jones, S., Harris, D.E., Katzer, F., Larke, N., Lord, A., Maser, P., McKellar, S., Mooney, P., Morton, F., Nene, V., O'Neil, S., Price, C., Quail, M.A., Rabbinowitsch, E., Rawlings, N.D., Rutter, S., Saunders, D., Seeger, K., Shah, T., Squares, R., Squares, S., Tivey, A., Walker, A.R., Woodward, J., Dobbelaere, D.A., Langsley, G., Rajandream, M.A., McKeever, D., Shiels, B., Tait, A., Barrell, B., Hall, N., 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science 309, 131-133.

Roy, S.W., Penny, D., 2006. Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. Genome Res. 16, 1270-1275.

Roy, S.W., Penny, D., 2007. Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. Mol. Biol. Evol. 24, 1926-1933.

Saeij, J.P., Coller, S., Boyle, J.P., Jerome, M.E., White, M.W., Boothroyd, J.C., 2007. *Toxoplasma* co-opts host gene expression by injection of a polymorphic kinase homologue. Nature 445, 324-327.

Sager, H., Brunschwiler, C., Jungi, T.W., 1998. Interferon production by *Theileria annulata*-transformed cell lines is restricted to the beta family. Parasite Immunol. 20, 175-182.

Sharp, P.M., Matassi, G., 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. 4, 851-860.

Shiels, B., Langsley, G., Weir, W., Pain, A., McKellar, S., Dobbelaere, D., 2006. Alteration of host cell phenotype by *Theileria annulata* and *Theileria parva*: mining for manipulators in the parasite genomes. Int. J. Parasitol. 36, 9-21.

Shiels, B.R., McKellar, S., Katzer, F., Lyons, K., Kinnaird, J., Ward, C., Wastling, J.M., Swan, D., 2004. A *Theileria annulata* DNA binding protein localized to the host cell nucleus alters the phenotype of a bovine macrophage cell line. Eukaryot. Cell 3, 495-505.

Swan, D.G., Phillips, K., Tait, A., Shiels, B.R., 1999. Evidence for localisation of a *Theileria* parasite AT hook DNA-binding protein to the nucleus of immortalised bovine host cells. Mol. Biochem. Parasitol. 101, 117-129.

Swan, D.G., Stern, R., McKellar, S., Phillips, K., Oura, C.A., Karagenc, T.I., Stadler, L., Shiels, B.R., 2001. Characterisation of a cluster of genes encoding *Theileria annulata* AT hook DNA-binding proteins and evidence for localisation to the host cell nucleus. J. Cell Sci. 114, 2747-2754.

Taylor, S., Barragan, A., Su, C., Fux, B., Fentress, S.J., Tang, K., Beatty, W.L., Hajj, H.E., Jerome, M., Behnke, M.S., White, M., Wootton, J.C., Sibley, L.D., 2006. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*. Science 314, 1776-1780.

Weir, W., Ben Miled, L., Karagenc, T., Katzer, F., Darghouth, M., Shiels, B., Tait, A., 2007. Genetic exchange and sub-structuring in *Theileria annulata* populations. Mol. Biochem. Parasitol. 154, 170-180.