Identification of candidate transmission-blocking antigen genes in Theileria annulata and related vector-borne apicomplexan parasites

Citation for published version:
Lempereur, L, Larcombe, SD, Durrani, Z, Karagenc, T, Bilgic, H, Bakirci, S, Hacilarlioglu, S, Kinnaird, J, Thompson, J, Weir, W & Shiels, B 2017, 'Identification of candidate transmission-blocking antigen genes in Theileria annulata and related vector-borne apicomplexan parasites' BMC Genomics, vol. 18. DOI: 10.1186/s12864-017-3788-1

Digital Object Identifier (DOI):
10.1186/s12864-017-3788-1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
BMC Genomics

Identification of candidate transmission-blocking antigen genes in *Theileria annulata* and related vector-borne apicomplexan parasites

Laetitia Lempereur[ab*], Stephen D Larcombe[a*], Zeeshan Durrani[ac], Tulin Karagenc[d], Huseyin Bilgic[d], Serkan Bakirci[d], Selin Hacilarlioglu[d], Jane Kinnaird[a], Joanne Thompson[e], William Weir[a], Brian Shiels[a+]

[*] equal contribution as first co-authors

[a] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary & Life Sciences, University of Glasgow, 464 Bearsden Road, Glasgow G61 1QH, Scotland, UK

[b] Current address: Laboratory of Parasitology and Parasitic Diseases, Department of Infectious and Parasitic Diseases, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

[c] Current address: School of Veterinary Science, University of Liverpool, Chester High Road, Neston, CH64 7TE, UK

[d] Adnan Menderes University, Faculty of Veterinary Medicine, Department of Parasitology, Batı Kampus, Işıklı, Aydın, Turkey

[e] Institute of Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, The King's Buildings, Edinburgh, EH9 3FL, UK

[+] Corresponding author

e-mail addresses:

Laetitia Lempereur: llempereur@ulg.ac.be; Stephen D Larcombe: Stephen.larcombe@glasgow.ac.uk , Zeeshan Durrani: Z.Durrani@liverpool.ac.uk: Tulin Karagenc: tulinkaragenc@yahoo.com; Huseyin Bilgic: huseyin_bilgic@yahoo.com , Serkan Bakirci: serkanbakirci@adu.edu.tr; Selin

Hacilarlioglu: selin-uner@hotmail.com; Jane Kinnaird: jane.kinnaird@glasgow.ac.uk; Joanne Thompson: Joanne.Thompson@ed.ac.uk; William Weir: willie.weir@glasgow.ac.uk; Brian Shiels: brian.shiels@glasgow.ac.uk

Abstract

Background

Vector-borne apicomplexan parasites are a major cause of mortality and morbidity to humans and livestock globally. The most important disease syndromes caused by these parasites are malaria, babesiosis and theileriosis. Strategies for control often target parasite stages in the mammalian host that cause disease, but this can result in reservoir infections that promote pathogen transmission and generate economic loss. Optimal control strategies should protect against clinical disease, block transmission and be applicable across related genera of parasites. We have used bioinformatics and transcriptomics to screen for transmission-blocking candidate antigens in the tick-borne apicomplexan parasite, *Theileria annulata*.

Results

A number of candidate antigen encoding genes were identified which included domains that are conserved across vector-borne Apicomplexa (*Babesia*, *Plasmodium* and *Theileria*), including the Pfs48/45 6-cys domain and a novel cysteine-rich domain. Expression profiling confirmed that selected candidate genes are expressed by life cycle stages within infected ticks. Additionally, putative B cell epitopes were identified in the *T. annulata* genes encoding the 6-cys and cysteine rich domains, in a gene encoding a putative papain-family cysteine peptidase, with similarity to the *Plasmodium* SERA family, and the gene encoding the *T. annulata* major merozoite/piroplasm surface antigen, *Tams1*.

Conclusions

Candidate genes were identified that encode proteins with similarity to known transmission blocking candidates in related parasites, while one is a novel candidate conserved across vector-borne apicomplexans and has a potential role in the sexual phase of the life cycle. The results indicate that a ‘One Health’ approach could be utilised to develop a transmission-blocking strategy effective against vector-borne apicomplexan parasites of animals and humans.

Keywords: *Theileria annulata*, *Plasmodium*, *Babesia*, bioinformatic screen, transmission-blocking vaccine, 6-Cys domain

Background

Tropical theileriosis is a lymphoproliferative disease of cattle that occurs from Southern Europe and North Africa in the west, through the Middle East, Central Asia and Indian sub-continent, to China in the east. The disease is caused by infection of bovines with the tick-borne apicomplexan parasite *Theileria annulata* and is a severe constraint to livestock productivity. Tropical theileriosis can show acute and chronic forms; with acute disease characterised by fever, weakness and emaciation, swelling of superficial lymph nodes, destruction of the lymphoid system and pulmonary oedema. Death from acute theileriosis is common in susceptible *Bos taurus* cattle and can occur within 21-28 days. Overt theileriosis has been a major problem in endemic regions when European cattle have been imported to improve livestock productivity. However, it is likely that the economic loss from animals undergoing chronic disease or showing no apparent clinical signs (carriers) is greater than that due to overt disease. This was demonstrated in a Tunisian study where up to 38 % of overall losses attributable to tropical theileriosis were associated with reduced milk production by carrier animals [1, 2]. Thus, to optimise economic output of cattle production in endemic regions, total control of theileriosis and related tick-borne disease (TBD) is required.

Current control measures include the use of acaricides, chemotherapy (primarily buparvaquone) and vaccination. Vaccination, with infected cell lines that develop attenuated virulence upon long-term culture, has been utilised in several countries [2, 3]. These vaccines can provide protection against clinical disease in the field but do not prevent establishment of carrier status. Thus, vaccination does not negate economic loss or the possibility of onward transmission from immunised carrier animals. In addition, for live vaccines there are potential risks of contamination with viral pathogens and reversion to virulence, and good quality control and a cold chain are required for effective delivery. Due to these disadvantages, plus recent reports of resistance to buparvaquone [4] and problems with continued use of acaricides (reviewed in [5]) there is a clear need for research into alternative, complementary control strategies.

An obvious strategy to control tropical theileriosis, and other TBD, is to prevent onward transmission of the pathogen by the tick vector. The efficacy of targeting ticks to block disease transmission is well known and has been validated by modelling studies, risk factor analysis and deployment of acaricides [6-8]. Use of acaricides, however, has an environmental impact and leads to selection of acaricide-resistant ticks [9]. The potential for anti-tick subunit vaccines to control tick infestation and decrease acaricide use has been demonstrated [10], with studies on the hidden gut antigen of *Boophilus microplus* (BM86) providing a paradigm model. Vaccination of cattle using the BM86 orthologue of *Hyalomma anatolicum anatolicum* (HAA86) showed that the tick gut antigen partially protected against homologous tick challenge and also reduced transmission of *Theileria annulata* [11]. In addition to targeting the tick, the potential of targeting surface antigens of the *Theileria* sporozoite and piroplasm stages to block transmission has been investigated. Antibodies against SPAG1 can effectively block invasion of the leukocyte by the sporozoite, while a response against the immunodominant Tams1 antigen has been implicated in blocking transmission of predominant genotypes [12, 13]. However, both these antigens show a degree of antigenic diversity in the parasite population that restricts their effectiveness as vaccine candidates [14-17]. This is particularly pertinent for Tams1 with identification of many allelic sequences, evidence of domain shuffling to generate molecular mosaics and the breakthrough of under-represented genotypes encoding variant Tams1 alleles when a carrier infection is transmitted through ticks [13, 14].

In order to circumvent antigenic diversity, proteins that perform a function that requires polypeptide domains to be invariant in the parasite population could be targeted. A potential advantage of selecting conserved protein domains is that they may be effective across a range of vector-borne diseases, by targeting processes or antigens common across related pathogens. One process of vector-borne Apicomplexans (*Babesia*, *Plasmodium* and *Theileria*) that could involve molecules conserved across genera is the sexual phase of the life cycle, which is obligatory for transmission of these parasites through their arthropod hosts. Proteins that function in the sexual phase and have potential to induce a transmission-blocking response against *Plasmodium spp.* have been identified. Surface antigens such as Pfs230, Pfs48/45, and Pfs25 are known to induce an immune response in vaccinated mammalian hosts that blocks transmission through the mosquito, thus demonstrating the feasibility of single or multi-subunit transmission blocking vaccines (TBVs) [18-22]. A considerable number of potential TBV candidates that perform functions required during the mosquito phase of the *Plasmodium* life-cycle have since been characterised (reviewed in [23]).

In the present study, a screen for parasite antigens with the potential to induce a transmission-blocking response against *T. annulata* was conducted. A combination of bioinformatic prediction and transcriptional expression profiling was used to obtain a panel of candidates, a number of which have homologues across genera of related vector-borne Apicomplexa. Analysis of the expression levels of four candidate genes in the tick vector, together with investigation of their predicted antigen diversity (*in silico*) provides evidence that development of transmission-blocking strategies which can operate across related vector-borne Apicomplexa may be possible.

Methods

Bioinformatic screening

A bioinformatic approach was used to identify *Theileria annulata* genes encoding proteins predicted to be located on the parasite surface using information representing 3772 genes contained in the genomic databases, GeneDB (http://www.genedb.org/Homepage/Tannulata) and EuPathDB (http://eupathdb.org). Genomic annotation data was downloaded using the 'List Download' feature of GeneDB. Candidate genes encoding putative surface antigens were selected on the basis of motifs predicted to be present on the encoded protein, namely a signal peptide, a GPI-anchor signal and/or one or more transmembrane domains. Database prediction for signal peptide (SignalP 2.0 HMM), GPI-anchor signal (DGPI v2.04) and transmembrane domains (TMHMM Server v2.0) were utilised using default settings. For *TA20855* and related homologues, sequences were also analysed using the SignalP 3.0 Server (http://www.cbs.dtu.dk/services/SignalP-3.0/) and SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP-4.1/). Integral membrane proteins with multiple predicted transmembrane domains were excluded. A subset of *Theileria annulata* genes that display elevated levels of mRNA expression from the macroschizont to the piroplasm stage of the life cycle in the vertebrate host were identified using a published microarray dataset [24, 25]. Hierarchical clustering of log2-transformed gene expression levels and profiles of gene expression values across stages (sporozoite to piroplasm) were performed using DNASTAR Array Star3 software, as described [25]. The NCBI database was BLAST searched (https://blast.ncbi.nlm.nih.gov/) to identify homologues of candidate genes in other vector-borne Apicomplexan parasites, namely *Plasmodium spp*., *Theileria spp*. and *Babesia spp*.

Revised annotation of *TA20855* and *TA19820*

Following alignment of homologs across genera for *T. annulata* genes *TA20855 and TA19820* it was observed that conservation of amino acid sequence observed for other apicomplexa was not obtained with sequence predicted for the *T. annulata* genes. Analysis of the gene DB entry sequence, however, showed both genes contained multiple introns and sequence with greater identity to the predicted amino acid sequence conserved across genera. An altered open reading frame was then identified and used to generate a revised amino acid sequence with greater conservation across genera. In order to verify that the revised predicted amino acid sequences were accurate, we used available next generation sequencing data. RNA-seq reads generated from sheep B-cells inoculated with *T annulata* stabilate (Ta Ankara, stabilate 89) for another experiment were kindly provided by Prof Ivan Morrison (Roslin Institute, University of Edinburgh). These RNA reads were of sufficient depth to provide coverage across the predicted *TA20855* and *TA19820* genes. Using the Bowtie 2 sequence aligner [24] RNA-seq reads were aligned to the predicted CDS of *TA20855* and *TA19820* provided on GeneDB. As expected, analysis of the created contigs revealed significant gaps in coverage, suggesting incorrect annotation and the presence or absence of introns.

The revised predicted amino acid sequences (designed to maximise orthology across Apicomplexa) were then aligned with the gDNA sequences (using Genewise Protein-nucleotide alignment software) to generate a new gene model and predicted CDS for both genes. The RNA seq reads were then mapped to the new predicted CDS sequences using Bowtie 2. The revised contigs showed much greater overlap between reads and coverage, including regions where incorrectly annotated introns and exons were responsible for the frame shifts in the original GeneDB gene models. Further revision to the predicted CDS was made to close gaps in RNA-seq coverage caused by other unidentified intron or exons, resulting in complete coverage and overlapping mRNA reads across the CDS (see Figures A and D in Additional file1). For *TA20855*, the final gene model results in 11 exons in contrast to 8 in the GeneDB model, while for *TA19820* the revised gene model results in the lengthening of 3 introns, and the inclusion of one more intron in contrast to the GeneDB entry (see Figures B and E in Additional file 1). The revised mRNA sequence for both genes is extremely similar to the reference genome sequence, with only a handful of SNPs.

qRT-PCR on selected candidate genes for a time course of T. annulata infected ticks

4100 ticks (*Hyalomma anatolicum anatolicum*) were fed on a calf infected with *T. annulata* Ankara sporozoite stabilate A10/BT (applied to the calves on Day 8 to Day 12 post-infection) with the parasitaemia peaking at 4 % on Day 14. Engorged ticks were collected (stored at 15 °C, until collection of all ticks post-detachment) and then incubated at 28 °C for 2, 6, 10 and 15 days, followed by freezing in RNAlater® (Thermo Fisher Scientific) at -80 °C. These time-points represent early events in gametocyte maturation (Day 2) together with gamete (Day 6 -10), zygote (Day 10) and kinete (Day 15) production, as reported previously [25, 26]. 400 frozen ticks for each time-point were crushed in liquid nitrogen and RNA extracted using TRIzol Reagent (Invitrogen) following the manufacturer's protocol. Four candidate genes (*TA10955*, *TA17050*, *TA03640* and *TA20855*) were selected for qRT-PCR based on bioinformatically predicted characteristics, microarray gene expression profile and detection of orthologues in other vector-borne Apicomplexa. Primers were designed (Additional file 2) and qRT-PCR was performed as described previously [27]. Briefly, 500 ng of total RNA from each sample was used to synthesise cDNA using the Affinity Script cDNA Synthesis Kit (Agilent Technologies) and Oligo-dT as primer.

1 µl cDNA for each sample was used for qRT-PCR, using the Brilliant III Ultra-fast SYBR®Green qPCR Master mix (Agilent technologies) and the Stratagene Mx3005P system. Comparative quantitative analysis of gene expression across time-points was performed using Stratagene MxPro Software, with RNA from a merozoite Day 8 culture used as the calibrator. HSP70 (*TA11610*) and HSP90 genes (*TA10720*) were utilised as controls for constitutive expression, based on their transcriptional profile through the life-cycle [27]. Differences in mean fold-change between time-points in candidate gene expression level were tested using Student's t-test; P-values obtained are denoted in the Results section and in Figure Legends.

Allelic dN/dS and epitope mapping for selected candidate genes

Analysis of allelic sequences generated from DNA samples from different *T. annulata* isolates from four different geographic origins was performed, with ratios of dN/dS computed to screen for evidence of diversifying positive selection for amino acid substitution on a codon-by-codon basis. The DNA samples were: *T. annulata* Ankara (Turkey), Hissar (India), 9A (Tunisia) and UmBanein24 (Sudan). PCR for genes of interest was performed on DNA from each of the four isolates and the resulting PCR amplicons were cloned and sequenced. Primers were specifically designed to amplify almost the entire length of *Tams1* (*TA17050*), putative papain-family cysteine protease (*TA10955*), and hypothetical protein *TA20855* (Additional file 2). *TA03640* was too large for the whole gene to be sequenced (>3000 bp), so for preliminary analysis two shorter fragments (~1500 bp and 2000 bp) were amplified, and five test sequences for each fragment generated. As this showed the second segment of the gene to be more polymorphic than the first, this region was chosen for further investigation of allelic polymorphism. Optimum annealing temperatures for each primer pair were determined (Additional file 6) and, to minimise the chance of PCR error in amplicons, *Pfu* Turbo DNA polymerase (Agilent Technologies) or KAPA HiFi (Kapa Biosystems,) polymerase was used in the PCR reaction, according to the manufacturer's guidelines. PCR products were cloned into pCR$^{®}$4Blunt-Topo vector (Invitrogen) and used to transform competent *E. coli*, using standard methodology. For *TA17050* and *TA10955*, twelve colonies from each isolate were selected and inserts sequenced in both directions (96 total sequences for each gene) by Genoscreen (Lille, France). For *TA20855* and *TA03640*, six colonies from each isolate were selected and sequenced in both directions (48 total sequences for each gene) by Eurofins (Berlin, Germany). The assembled sequences were translated and aligned to the GeneDB reference amino acid sequences (GenBank accession n° XP_953719, XP_953243, XP_954368) using CLC Genomics Workbench software and polymorphic sites identified. The datasets of allelic sequences were then used to estimate the ratio of non-synonymous to synonymous base-pair substitutions (dN/dS) for each codon in each gene and for the entirety of the selected gene or region using the SLAC algorithm of the online Datamonkey program (http://www.datamonkey.org [28]). The SLAC method is a conservative method for calculation of dN/dS that prevents overestimation of positive selection [29]. Finally, we used the Bepipred linear B-cell epitope prediction tool (http://tools.immuneepitope.org/tools/bcell/iedb_input) [30] to predict areas of each gene that could form B-cell epitopes. Data from both types of analysis were then overlaid to visualise any regions for each candidate gene where evidence for selection of amino acid substitution and prediction of a B cell epitope overlapped.

Results

Bioinformatic and transcriptomic profile analysis identifies T. annulata transmission-blocking candidate genes

To screen for *T. annulata* candidate genes encoding proteins that may be expressed by life cycle stages present in the tick vector a combined genomic and transcriptomic approach was taken. A screen of available genomic data was used to identify genes encoding proteins with a predicted signal peptide domain together with a GPI anchor domain, resulting in a list of 44 genes. Seven genes did not have direct orthologues in the closely related *Theileria parva* and were removed from the list. A further seven genes encoding proteins with multiple transmembrane domains were also removed, as likely to be integral membrane proteins, and so potentially less suitable as transmission blocking candidate antigens, leaving 30 candidate surface protein encoding genes. Microarray-derived transcriptomic data across all bovine life-cycle stages together with the tick-derived sporozoite stage was then analysed for these candidate genes [27]. From this, a subset of genes was selected which displayed an expression pattern that indicated rising mRNA levels from macroschizont through to piroplasm, the stage that is taken-up by the tick vector. Genes were selected on the basis of an absolute fold-change of greater than 2 between the macroschizont and merozoite and/or piroplasm stages. This resulted in a subset of 13 candidate genes (Table 1).

Expression profiles were constructed across all stages, for which data was available, with candidate genes grouped into three profile types (see Figure 1). In the first profile (A): genes *TA13810*, *TA17050* and *TA20855* showed elevation of expression from macroschizont through to the merozoite/piroplasm and the level of expression in the sporozoite was similar to that of the macroschizont (<2 fold difference). In the second profile (B): genes *TA02580*, *TA03640*, *TA03755*, *TA16565*, *TA17220*, showed elevation of expression through to the merozoite/piroplasm stages and had a significantly higher level of expression (>2 fold) in the sporozoite relative to the macroschizont stage, and the expression level in the sporozoite was higher than that for the merozoite/piroplasm. In the third profile (C): genes *TA02585*, *TA16005*, *TA10955* and *TA13825* showed expression levels that were higher in sporozoite than in macroschizont and either comparable between sporozoite and piroplasm or lower in the sporozoite. Within these profiles two genes previously considered as transmission-blocking candidates were identified. *Tams1* (*TA17050*) showed expression consistent with profile A, while *SPAG1* (*TA03755*) showed expression consistent with profile B. It was concluded that these profiles indicate the potential for the gene to be expressed, either transiently in the tick following a blood meal, or at an elevated level that is coincident with the development of stages within the tick vector.

Identification of candidate genes showing domain conservation across vector-borne Apicomplexa

To determine if any of the candidate genes are predicted to possess domains that perform a conserved function, their entries in GeneDB were examined and BLAST analysis for homologues in other Apicomplexa was carried out. *Tams1* (TA17050) and *SPAG1* (*TA03755*) have been characterised extensively with orthologues identified across the *Theileria* genus, they have no known domains that show conservation across the vector-borne Apicomplexa.

TA02580 and *TA02585* encode putative surface proteins of unknown function with respective orthologues only identified in *T. parva*. *TA13810* was identified as the direct orthologue of the gene encoding the ts-chitose type 23 kDa piroplasm surface protein of *T. sergenti* [31] and is conserved across bovine *Theileria spp.*; an orthologue was not identified for other vector-borne Apicomplexa. Similarly, orthologues of *TA13825* were only identified in *Theileria ssp.* (*T. parva*, *T. orientalis/buffeli* and *T. equi*) and showed similarity to the 23 kDa piroplasm surface protein. *TA16005* encodes a protein of unknown function that is also restricted to *Theileria* species. *TA17220* has a probable orthologue in *T. parva* and shows similarity to an uncharacterised predicted protein in *T. orientalis*. *TA16565* is annotated as an uncharacterised surface protein with orthologues in both *Theileria* and *Babesia* genera (E-value-1.5E-17, B. bovis; - 9.5E-19, B. bigemina; 3.2E-25, B microti).

TA10955 is annotated in GeneDB as encoding a putative papain-family (clan CA) cysteine protease (Pfam: PF00112, E-value=1.86E-12) with a signal peptide and GPI anchor. BLAST analysis identified similarity (30 % identity, 49 % similarity) to the Serine repeat antigen 5 of *Plasmodium falciparum* that covers the predicted peptidase domain (218 - 476) of TA10955. Conservation of the domain (see Figure 2A) was found in predicted proteins of related *Theileria* species (*T. orientalis* and *T. parva*) as previously reported [32], but not in *Babesia* or *T. equi*.

Gene *TA03640* is annotated as encoding a hypothetical protein with a signal peptide and GPI anchor. It is also annotated as encoding an s48_45 domain between aa 1020-1135 (Pfam: 07422, E-value-1.2E-17) found in the 6-cys family of *Plasmodium* surface proteins (e.g. Pfs 48/45 and Pfs 230) that play an important role in gamete fertilisation in *Plasmodium* [33, 34]. The domain contains 6 conserved cysteines that form 3 disulphide bridges necessary for correct protein folding. The s48/45 domain is conserved across the vector-borne Apicomplexa with orthologues present in *Theileria* and *Babesia* species, as well as *Plasmodium*. The alignment represented in Figure 2B shows strong positional conservation of the 6 cysteines of the *Theileria* domain with orthologues in *Plasmodium* and *Babesia*.

TA20855 is annotated in Gene DB as encoding a hypothetical surface protein of 297 aa with similarity to *Plasmodium* hypothetical proteins. Clear homologues with significant similarity (>50 %) were identified by BLAST across the Apicomplexa (*Babesia*, *Plasmodium*, *Toxoplasma*, *Hammondia*), with the highest level of similarity spanning a region containing conserved cysteine residues. However, based on identity of predicted amino acid sequences across other genera, compared to that identified for the *Theileria* orthologues, it was concluded that the original annotation of intron exon junctions in *TA20855* on GeneDB predicted an incorrect open reading frame, with the *TA20855* sequence diverging from those of related genera at aa 255. A revised gene model (based on homology of predicted aa sequence across genera) encoding a protein of 289 aa was then validated using available RNA seq data, with complete coverage of the revised polypeptide coding sequence obtained (see Additional file 1). Using the revised gene model, homology over a region spanning aa 128 – 282 (58 % similarity; minimum E-value 2E-18) was found across genera of Apicomplexa, and *Vitrella brassicaformis*, a chromerid that evolved from a common ancestor shared with the Apicomplexa [35]. This region of homology contains 8 positional conserved cysteine residues and spans at least one predicted transmembrane (TM) helix, with a second more C-terminal helix predicted in some instances (depending on the sequence modelled or the algorithm used; see Figure 2C and Additional file 1. Both these predictions, plus the prediction of a cleaved signal peptide (*T. annulata*; *B. bovis*, *B. bigemina*, *Vitrella brassica* (using both SignalP 3.0 and 4.1) and *T. gondii* (SignalP

3.0 but not 4.1)), indicate that the majority of the encoded polypeptide (a helical rich region) is to the extracellular side of the membrane. For *Plasmodium* polypeptides, while a cleaved signal peptide is predicted by SignalP 3.0, an alternative model with the helical rich region on the cytoplasmic face is also indicated, as a transmembrane helix is also denoted within the putative signal peptide region in their EuPathDB entries. Thus, gene *TA20855* is likely to encode a membrane protein that is conserved across related vector borne genera and was present in a common ancestor of the apicomplexans and chromerids.

Elevated expression of candidate genes in tick stages of T. annulata

To assess potential expression of selected candidate genes in tick stages of *T. annulata*, qRT-PCR was performed on RNA representing a developmental time-course after engorgement of ticks on a piroplasm-infected animal. Four genes were selected representing the A (*TA17050* and *TA20855*), B (*TA03640*) and C (*TA10955*) microarray expression profiles. These genes include a *Theileria*-specific candidate (*TA17050*), a candidate that showed homology with *Plasmodium* proteins (*TA10955*) and two candidates (*TA03640* and *TA20855*) with homologues present in vector-borne Apicomplexa.

As illustrated in Figure 3A, the expression profile for *TA17050* (*Tams1*) showed a drop in expression at day 2, relative to the merozoite stage calibrator RNA (2.1-fold, absolute), which continued as the infected tick time-course progressed to Day 15 (39-fold reduction). *TA10955* (the putative papain-family cysteine protease gene) showed a decrease (>3-fold, absolute) in expression at Day 2 post-detachment of ticks relative to merozoite RNA (see Figure 3B). However, in contrast to *Tams1,* from day 2 onwards, expression of *TA10955* increased as the tick time-course progressed, with a marked significant ($p < 0.001$) elevation at Day 15 (> 7000 fold absolute at Day 15, relative to merozoite RNA).

For *TA03640*, significant elevated expression was not detected at the early time-points. By Day 10, however, expression levels were increased significantly relative to merozoite and Day 2 (>13 fold absolute, $p < 0.0001$), and this was sustained at Day 15 (Figure 3 C). A related expression profile was obtained for *TA20855* with expression significantly elevated at Day 6, relative to Day 2 ($p = 0.002$), and a further increase (>3-fold absolute, $p < 0.001$) at Day 10 relative to Day 6 (Figure 3D).

However, unlike *TA03640*, a significant fall ($p < 0.001$) in expression between Day 10 and Day 15 occurred (>10-fold, absolute decrease) to a level below that of the Day 6 time-point (>3-fold, absolute decrease). To compare expression profiles for the *Plasmodium and Toxoplasma* homologues of *TA20855*, data available in EuPathDB was mined. This demonstrated that for all *Plasmodium* homologues for which data is available, RNA is up-regulated in late stage (V) gametocytes, indicating a putative role in transmission via the mosquito vector (Additional file 2), while in *Toxoplasma* the highest level of expression was associated with unsporulated oocysts.

Assessment of dN/dS and in silico prediction of B cell epitopes of transmission-blocking candidate genes

Genes encoding antigens exposed to a protective immune response often display an elevated ratio of non-synonymous (dN) nucleotide substitution to synonymous substitution (dS) across allelic sequences [36, 37]. In contrast, genes encoding proteins specific to vector stages and not exposed to an acquired protective immune response may show more limited levels of selection for amino acid substitution [38]. To assess whether the putative proteins encoded by transmission-blocking candidate genes may be exposed to the immune response or act as hidden antigens, the level of dN/dS was computed for three candidate genes with evidence of elevated expression in tick stages. This was performed in comparison to the *Tams1* gene, as the level of dN/dS has been found to be relatively high among *Tams1* alleles [14, 16]. Allelic sequences were generated for all four selected genes from DNA representing a panel of parasite isolates: *T. annulata* Ankara (Turkey), Hissar (India), 9A (Tunisia) and UmBanein24 (Sudan). For each gene a minimum of 48 sequences were obtained and distinct consensus sequences selected. The dN/dS ratio was then computed as: 0.48 for *Tams1* (*TA17050*) with six significantly positive selection sites at $p < 0.1$; 0.13 for TA10955 with three significantly positive selection sites at $p < 0.1$; 0.19 for *TA03640* with no positively selected sites at $p < 0.1$; and 0.31 for *TA20855* with no significantly positively selected sites at $p < 0.1$. Thus, as expected, *Tams1* (*TA17050*) was shown to be the gene with strongest evidence for selection of amino acid substitutions. In contrast, the overall dN/dS ratio of *TA10955*, *TA03640* and *TA20855* was lower and few (*TA10955*) or no statistically significant positively selected sites were identified. However, visualisation of dN/dS plots (Figure 4) revealed a degree of clustering of codons where dN/dS values were positive; this indicated that non-synonymous amino acid substitutions were tolerated, although there was insufficient power to determine these as statistically significant.

Amino acid substitutions that are positively selected to allow evasion from immune detection could be expected to coincide with the position of antigenic epitopes of *Theileria* proteins, as demonstrated for the T cell antigen gene, TA9 [37]. We used B-cell antibody epitope prediction software to identify regions of each protein where putative antibody epitopes could be detected. Using the Bepipred linear epitope prediction algorithm, individual amino acid residues were denoted as being within or outside predicted B-cell epitopes and this data superimposed onto the dN/dS plots. Preliminary inspection suggested that for at least some of the candidate genes, overlap exists between regions harbouring amino acid substitutions and predicted B cell epitopes (see Figure 4). To test overlaps for statistical significance, every residue was classified for predicted epitope (yes or no) and evidence of positive selection (i.e. positive dN/dS values: yes or no) and a chi-square test performed. For candidate genes (*TA10955* and *TA20855*) there was a robust relationship between regions of amino acid substitution and regions of predicted epitopes: thus, sites with positive dN/dS scores, though rare, were significantly more likely to occur in regions of predicted epitopes than in non-epitope regions (*TA10955* $X^2 = 6.53$ $p = 0.011$; *TA20855* $n = 243$ $X^2 = 938$ $p = 0.002$). Taken together, the results suggest that these two candidate genes encode polypeptides with putative B-cell epitopes that exhibit weak, but detectable, evidence for selection of amino acid substitution.

Discussion

The primary aim of this study was to use a bioinformatic approach to identify candidate genes encoding proteins with the potential to induce an immune response that could block transmission of *Theileria annulata* by the tick vector. Moreover, given that sexual reproduction is likely to have been retained by all vector-borne Apicomplexa [39], a secondary aim was to identify candidate genes that show a degree of conservation across *Theileria* species and related genera, particularly *Babesia*, as the two genera can be endemic over the same geographical region [40]. Two types of transmission-blocking candidate antigen were predicted: firstly (type 1), surface antigens required for the early phase of infection in the tick may be expected to be present in the bovine host and exposed to its acquired immune response, hence these antigens were expected to display a degree of antigenic diversity; secondly (type 2), surface proteins exclusive to stages present within the tick that perform an important biological function, such as gamete fertilisation, and may possess conserved epitopes that could induce a transmission-blocking antibody response if used as an antigen.

Based on results of our screen we identified twelve candidate genes, some of which possessed characteristics that allow placement into either type 1 or type 2 antigens. Thus, genes whose microarray expression level is elevated in the piroplasm but lower in the sporozoite stage are more likely to be expressed as proteins at the merozoite/piroplasm stage and may be present only in the initial phase of infection in the tick vector. This premise is supported by the observation that this group includes the genes, *TA17050* and *TA13810*, which encode the known major merozoite/piroplasm surface antigen, Tams1, and the 23 kDa piroplasm surface antigen. Expression in the tick was determined for *Tams1* where the RNA level was shown to fall within 2 days (at 28 °C post-detachment) and continued to fall over the remainder of the tick time-course. Thus, it can be predicted that synthesis of Tams1 protein (and by extrapolation possibly the 23 kDa piroplasm surface antigen) is significantly reduced (or absent) following generation of gamete forms (Day 6 - 10). This does not preclude a role for Tams1 as a transmission-blocking candidate, as piroplasms may persist for days within the tick; the protein may be stable and the antigen has been detected in infected tick protein extracts (Mohamed Dargouth, personal communication).

The results for *TA20855* show that it would be unwise to predict an expression profile for tick stages based solely on the available microarray data. TA20855 shows a similar microarray profile to that of *Tams1* (*TA17050*) but we have shown by qRT-PCR analysis that peak expression of *TA20855* does not occur until around Day 10 post-detachment, a time-point associated with gamete fertilisation and production of zygotes [25]. The rapid fall in expression at the Day 15 time-point suggests a transient role prior to the production of kinetes, with a logical prediction being that the encoded protein is specific to gametes and perhaps performs a role in fertilisation or zygote development, although a role post Day 10 cannot be totally discounted if the protein is highly stable. BLAST analysis revealed a region of considerable identity, particularly over a predicted 8-cysteine structural domain, with genes encoding predicted membrane proteins in other Apicomplexa and in the chromerid, *Vitrella*. Homologues in *Plasmodium* show a transient peak in late-stage gametocytes (Additional file 3), while differential expression in *Toxoplasma* shows elevated expression associated with the unsporulated oocyst. The results suggest that this gene could be a remnant of the ancestral machinery of apicomplexan sexual reproduction. Based on the gene model it is likely to be an integral membrane protein, but with a significant proportion predicted to be extracellular. The region of greatest amino acid identity across homologues contains 8 spatially conserved cysteines and spans the region(s) predicted to act as a transmembrane helix. TM helices with conserved patterns of residues are unusual and indicate potential functional significance [41]. This is supported by evidence for conservation of amino acid substitution across *T. annulata* alleles in the region of the molecule predicted to be within the cell membrane or cytoplasm (see Fig 4D). The function on the molecule can only be speculated upon at present, but the two most likely possibilities are as a ligand or a structural surface molecule that protects the parasite from the extracellular environment.

Genes that display an elevated level of RNA expression associated with the sporozoite stage may encode surface proteins whose function is primarily required after gamete fertilisation/zygote production. The gene (*TA03755*) encoding the major sporozoite surface antigen SPAG-1 [42] and a gene encoding a putative papain cysteine protease were placed in this category. *TA10955* was found to display peak expression at Day 15 of the tick time-course, indicating that the encoded protein may not be present until the later part of the life cycle in the tick. The predicted protein shows strongest similarity to the serine repeat antigen family (SERA) of *P. falciparum,* identified as important asexual blood-stage antigens (reviewed by [43]). The *Theileria* SERA represents a phylogenetic out-group to *Plasmodium* SERAs [32], with similarity over the peptidase domain of the predicted protein, but not the antigenic N-terminal domain identified for *Plasmodium* SERA5. Members of the *Plasmodium* SERA family function in merozoite egress, and have been implicated in sporozoite egress from the oocyst within the mosquito, providing a potential target for transmission blocking strategies [32, 44]. Based on its RNA expression pattern, the *Theileria* protein may function in a similar manner, promoting release of kinetes or sporozoites from infected tick cells. Whether this involves a surface associated location or secretion of the proteinase into the host cell environment would require validation.

The *TA03640* gene has an expression profile that is similar to *TA20855* but does not show a significant drop at Day 15. It is expressed at a higher level in the sporozoite relative to the merozoite/piroplasm, suggesting that production of the protein occurs within the tick. *TA03640* contains a pfs48_45 domain present in members of the 6-cys family in *Plasmodium*, including the gametocyte and gamete surface proteins Pfs48/45 and Pfs230 [33, 34]. Expression of a 6-cys encoding gene in *B. bovis* has been reported in merozoites [45]. However, a more recent study demonstrated low level expression in blood stages and elevated expression of at least seven out of ten *B. bovis* 6-cys genes by stages within the tick vector [46]. Thus, the elevated expression at the Day 10 and 15 time-points post-tick detachment suggests that *Theileria TA03460* may play a role in mediating transmission, as proposed for related vector-borne Apicomplexa. Based on demonstration that antibody responses generated against Plasmodium P48/45 and P230 can block transmission [34,[47] 6-cys domain surface antigens provide a target for development of transmission blocking vaccines against *Theileria* as well as *Babesia* [48].

The *in silico* analysis performed in this study indicated that the tested genes encode predicted B cell epitopes indicating that they have antigenic properties. In general, and relative to the merozoite/piroplasm major antigen gene *Tams1*, amino acid diversity is limited and no positively selected amino acid substitution sites were predicted for *TA03640* or *TA20855*. This could be taken as evidence that they may operate as “hidden antigens” and provide a target that shows conservation across species isolates, as proposed for the *Babesia* 6-cys candidate genes [48]. However, for

TA20855 (and *TA10955*) predicted epitopes coincided with certain positions where there is evidence for allelic amino acid substitution, and these were most evident in the region of the molecule modelled to be on the extracellular side of the cell membrane (i.e. N-terminal of residue 217). Evidence for selection of amino acid substitution has been reported for transmission-blocking proteins in *Plasmodium*, with the conclusion that this is driven by a protective immune response and/or mating interactions [34]. The former would require exposure to an acquired immune response, which has been shown to occur for *Plasmodium* transmission blocking antigens ([34] and references therein). Although our data indicates mRNA expression peaks in stages found within the tick, further work is required to assess whether polypeptides encoded by *TA10955*, *TA03640* and *TA20855* are hidden from or exposed to the immune response of the bovine host.

Tams1 (*TA17050*) alleles showed the highest dN/dS ratio with six significant positive selection sites: however, areas of amino acid substitution did not show strong co-localisation with predicted B cell epitopes. One possible explanation is that divergent epitopes for this surface antigen are thought to be highly conformational dependent, are sensitive to mild periodate treatment and may not have been predicted by the algorithm. In addition, epitopes that are internal to the folded molecule and are not exposed to a protective immune response are likely [48]. Thus, to be fully effective as a vaccine candidate an antigen profile encompassing a wide range of divergent epitopes would need to be generated.

Based on the results of this study, at least two genes (*TA03640*, *TA20855*) demonstrate that transmission-blocking candidates can show a degree of conservation across related genera (*Babesia*, *Theileria*, *Plasmodium*). This implies that additional candidates may exist, as several other classes of protein are known to play a functional role in transmission of *Plasmodium*. Indeed, a preliminary screen for *Theileria* orthologues of *Plasmodium* transmission-blocking candidate genes and analysis (using criteria defined in methods) of the expression profile in *T. annulata* yields several other candidates (see Additional files 4, 5 and 6) including: a second *Theileria* 6-cys gene (*TA14250*); a gene (*TA09115*) with orthology to genes encoding the HAP2 protein that has been proposed to function as a gamete membrane fusogen in *Plasmodium* and many other protists [49-51]; a gene (*TA19820*) encoding a domain with orthology to the CPW-WPC domain encoded by surface proteins associated with

Plasmodium transmission stages including the developing ookinete [52]. Further candidates are likely to be identified with a genome wide screen comparing bovine to tick stage transcriptome data.

Conclusions

A bioinformatics screen has identified candidate genes encoding proteins with characteristics that allow prediction they have potential to block transmission of *Theileria* parasites. Given the economic loss associated with sub-clinical infection of *T. annulata* and the role of carrier animals in generating new clinical outbreaks, we believe further testing of candidates using a multi-antigen approach, possibly combined with an anti-tick component [53, 54], is warranted. Since a degree of conservation across vector borne Apicomplexa clearly exists for genes that promote transmission through the arthropod, it should be possible to consider development of generic strategies that are effective against this important group of pathogens. Progress towards this goal will depend on funded vaccine trials, these may be expedited by using smaller animal models to test candidates conserved across Piroplasmida.

Declarations

Ethics approval and consent to participate

The only animal experiments in the study were performed in Turkey. The experimental protocols performed in the study were assessed and approved by the Adnan Menderes University Animal Experiment Ethic Committee (dated 26/08/2011), in accordance with decision number B.30.2.ADÜ.0.00.00.00/050.04/2011/058. The principle of the Three Rs was applied: no in vitro system exists that could replace the requirement of animals, only two animals were used and this was the minimum for generation of the required material. The animals used in the experiment were privately owned in Turkey, and the owner signed the relevant client-owner consent paperwork: a translated version of the report can be provided on request.

Availability of data and material

The allelic sequences of candidate genes generated and/or analysed during the current study are available in the NCBI repository with the following accession numbers: TA03640 KX980978 - KX98100; TA10955 KX981002 - KX981024;TA17050 KX981025 - KX981042; TA20855 KX981043 - KX981056. Other data from this study was mined from publicly available sources named throughout the text, and the rest is available in the manuscript or supplementary files.

Competing Interests

None

Funding

Funding for the project was provided through: the Farmed Animal Diseases and Health (FADH) initiative BBSRC (http://www.bbsrc.ac.uk): grant BB/L004739/1 and the EU (http://europa.eu/) who funded the allelic sequencing, qPCR, manuscript writing and bioinformatics analyses; POSTICK ITN (Post-graduate training network for capacity building to control ticks and tick-borne diseases) FP7-PEOPLE-ITN program (Grant No. 238511) who funded the allelic sequencing, and tick time-course; TUBITAK (TUBITAK-110O718) who funded the tick-time course and animal studies.

Consent for Publication

Not applicable

Acknowledgements

Thank you: Marta Pieszko for advice on qRT-PCR of differentially expressed *Theileria* genes and Andy Tait for discussion on screening strategy.

Author Contributions

LL and SL experimental design, generation of experimental data, data analysis and manuscript production; BRS study conception, experimental design, data analysis and manuscript production; WW and JT, bioinformatic analysis of microarray and genomic data, figure production, manuscript editing; JK and ZD, qRT-PCR experimental design and data analysis, editing of manuscript; TK, HB, SB, SH generation of *Theileria annulata* infected tick time course and maintenance of *Hyalomma* tick colony.

List of Abbreviations

TBD: Tick Borne Disease

TBV: Transmission Blocking Vaccine

GPI: Glycosylphosphatidylinositol

CDS: coding DNA sequence

TM: Transmembrane

dN/dS: ratio of non-synonomous (dN) to synonomous (dS) nucleotide substitutions

References

1. Gharbi M, Sassi L, Dorchies P, Darghouth MA. Infection of calves with *Theileria annulata* in Tunisia: Economic analysis and evaluation of the potential benefit of vaccination. Vet Parasitol. 2006; 137:231-241.
2. Gharbi M, Touay A, Khayeche M, Laarif J, Jedidi M, Sassi L, Darghouth MA. Ranking control options for tropical theileriosis in at-risk dairy cattle in Tunisia, using benefit-cost analysis. Revue Scientifique et Technique-OIE 2011;30:763.
3. Rasulov I, Fish L, Shkap V. Vaccination of cattle against tropical theileriosis in Uzbekistan using autochthonous live vaccine. Vaccine. 2008;26:G14-G16.
4. Mhadhbi M, Naouach A, Boumiza A, Chaabani MF, Ben Abderazzak S, Darghouth MA. *In vivo* evidence for the resistance of *Theileria annulata* to buparvaquone. Vet Parasitol. 2010; 169:241-247.
5. Benelli G, Pavela R, Canale A, Mehlhorn H. Tick repellents and acaricides of botanical origin: a green roadmap to control tick-borne diseases? Parasitol Res. 2016;1-16.
6. Graham OH, Hourrigan JL. Review article: Eradication programs for the arthropod parasites of livestock. J Med Entomol. 1977;13:629-658.
7. Jirapattharasate C, Moumouni PFA, Cao S, Iguchi A, Liu M, Wang G, Zhou M, Vudriko P, Changbunjong T, Sungpradit S. Molecular epidemiology of bovine *Babesia* spp. and *Theileria orientalis* parasites in beef cattle from northern and northeastern Thailand. Parasitol Int. 2016;65:62-69.
8. Sutton AJ, Karagenc T, Bakirci S, Sarali H, Pekel G, Medley GF. Modelling the transmission dynamics of *Theileria annulat*a: model structure and validation for the Turkish context. Parasitol. 2012;139:441-453.
9. George JE, Pound JM, Davey RB. 2004. Chemical control of ticks on cattle and the resistance of these parasites to acaricides. Parasitol 129:S353-S366.
10. Merino O, Alberdi P, de la Lastra JMP, de la Fuente J. Tick vaccines and the control of tick-borne pathogens. The biology and ecology of ticks shape the potential for the transmission of zoonotic pathogens Front Cell Infect Microbiol. 2015;93.
11. Jeyabal L, Kumar B, Ray D, Azahahianambi P, Ghosh S. Vaccine potential of recombinant antigens of *Theileria annulata* and *Hyalomma anatolicum anatolicum* against vector and parasite. Vet Parasitol. 2012;188:231-238.
12. Boulter NR, Brown CG, Kirvar E, Glass E, Campbell J, Morzaria S, Nene V, Musoke A, d'Oliveira C, Gubbels MJ, Jongejan F, Hall FR. Different vaccine strategies used to protect against *Theileria annulata*. Ann NY Acad Sci. 1998;849:234-246.
13. Gubbels MJ, Katzer F, Shiels BR, Jongejan F. Study of *Theileria annulata* population structure during bovine infection and following transmission to ticks. Parasitol. 2001;123:553-561.
14. Gubbels MJ, Katzer F, Hide G, Jongejan F, Shiels BR. Generation of a mosaic pattern of diversity in the major merozoite-piroplasm surface antigen of *Theileria annulata*. Mol Biochem Parasitol. 2000;110:23-32.
15. Katzer F, Carrington M, Knight P, Williamson S, Tait A, Morrison IW, Hall R. Polymorphism of *SPAG-1*, a candidate antigen for inclusion in a sub-unit vaccine against *Theileria annulata*. Mol Biochem Parasitol. 1994;67:1-10.

16. Katzer F, McKellar S, Ben Miled L, d'Oliveira C, Shiels B. Selection for antigenic diversity of *Tams1*, the major merozoite antigen of *Theileria annulata*. Ann NY Acad Sci. 1998;849:96-108.

17. Shiels BR, d'Oliveira C, McKellar S, Ben Miled L, Kawazu S, Hide G. Selection of diversity at putative glycosylation sites in the immunodominant merozoite/piroplasm surface antigen of *Theileria* parasites. Mol Biochem Parasitol 1995;72:149-162.

18. Carter R, Stowers A. Current developments in malaria transmission-blocking vaccines. Expert Opin Biol Ther. 2001;1:619-628.

19. Pradel G. Proteins of the malaria parasite sexual stages: expression, function and potential for transmission blocking strategies. Parasitol. 2007;134:1911-1929.

20. Kaslow DC. Transmission-blocking vaccines, In:Malaria Immunology, vol 80. Karger Publishers. p 287-307.

21. Carter R, Mendis KN, Miller LH, Molineaux L, Saul A. Malaria transmission-blocking vaccines—how can their development be supported? Nat Med. 2000;6: 241-244

22. Carter R. Transmission blocking malaria vaccines. Vaccine. 2001;19:2309-2314.

23. Wu Y, Sinden RE, Churcher TS, Tsuboi T, Yusibov V. Chapter Three-Development of Malaria Transmission-Blocking Vaccines: From Concept to Product. Adv Parasitol. 2015;89:109-152.

24. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat. meth. 2012; 9:357-359.

25. Pieszko M, Weir W, Goodhead I, Kinnaird J, Shiels B. ApiAP2 Factors as Candidate Regulators of Stochastic Commitment to Merozoite Production in *Theileria annulata*. PLoS Negl Trop Dis. 2015;9:e0003933.

26. Mehlhorn H, Schein E. The piroplasms: life cycle and sexual stages. Adv Parasitol. 1984;23:37-103.

27. Walker AR, Fletcher JD, McKellar SB, Bell LJ, Brown CG. The maintenance and survival of *Theileria annulata* in colonies of *Hyalomma anatolicum anatolicum*. Ann Trop Med Parasitol. 1985;79:199-209.

28. Pond SL, Frost SD. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2015;21:2531-2533.

29. Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2015;22:1208-1222.

30. Larsen JEP, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res. 2006;2:1.

31. Sako Y, Asada M, Kubota S, Sugimoto C, Onuma M. Molecular cloning and characterisation of 23-kDa piroplasm surface proteins of *Theileria sergenti* and *Theileria buffeli*. Int J Parasitol. 1999;29:593-599.

32. Arisue N, Hirai M, Arai M, Matsuoka H, Horii T. Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*. J Mol Evol. 2007;65:82-91.

33. van Dijk MR, Janse CJ, Thompson J, Waters AP, Braks JAM, Dodemont HJ, Stunnenberg HG, van Gemert G-J, Sauerwein RW, Eling W. A central role for P48/45 in malaria parasite male gamete fertility. Cell. 2001;104:153-164.

34. Van Dijk MR, Van Schaijk BCL, Khan SM, Van Dooren MW, Ramesar J, Kaczanowski S, van Gemert G-J, Kroeze H, Stunnenberg HG, Eling WM.

Three members of the 6-cys protein family of *Plasmodium* play a role in gamete fertility. PLoS Pathog. 2010;6:e1000853.

35. Moore RB, Oborník M, Janouškovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K. A photosynthetic alveolate closely related to apicomplexan parasites. Nature. 2008;451:959-963.

36. Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. Mol Biol Evol. 1996;13:685-690.

37. MacHugh ND, Weir W, Burrells A, Lizundia R, Graham SP, Taracha EL, Shiels BR, Langsley G, Morrison WI. Extensive polymorphism and evidence of immune selection in a highly dominant antigen recognised by bovine CD8 T cells specific for *Theileria annulata*. Infect Immun. 2011;79: 2059-2069

38. Kang J-M, Ju H-L, Moon S-U, Cho P-Y, Bahk Y-Y, Sohn W-M, Park Y-K, Cha SH, Kim T-S, Na B-K. Limited sequence polymorphisms of four transmission-blocking vaccine candidate antigens in *Plasmodium vivax* Korean isolates. Malar J. 2013;12:1.

39. Weedall GD, Hall N. Sexual reproduction and genetic exchange in parasitic protists. Parasitol. 2015;142:S120-S127.

40. Simuunza M, Weir W, Courcier E, Tait A, Shiels B. Epidemiological analysis of tick-borne diseases in Zambia. Vet Parasitol. 2011;175:331-342.

41. Venancio TM, Aravind L. CYSTM, a novel cysteine-rich transmembrane module with a role in stress tolerance across eukaryotes. Bioinformatics. 2010;26:149-152.

42. Williamson S, Tait A, Brown D, Walker A, Beck P, Shiels B, Fletcher J, Hall R. *Theileria annulata* sporozoite surface antigen expressed in *Escherichia coli* elicits neutralizing antibody. Proc Natl Acad Sci USA. 1989;86:4639-4643.

43. Palacpac NMQ, Arisue N, Tougan T, Ishii KJ, Horii T. *Plasmodium falciparum* serine repeat antigen 5 (SE36) as a malaria vaccine candidate. Vaccine. 2011; 29:5837-5845.

44. Aly ASI, Matuschewski K. A malarial cysteine protease is necessary for *Plasmodium* sporozoite egress from oocysts. J Exp Med. 2005;202:225-230.

45. Silva MG, Ueti MW, Norimine J, Florin-Christensen M, Bastos RG, Goff WL, Brown WC, Oliva A, Suarez CE. *Babesia bovis* expresses Bbo-6cys-E, a member of a novel gene family that is homologous to the 6-cys family of *Plasmodium*. Parasitol Int. 2011;60:13-18.

46. Alzan HF, Lau AOT, Knowles DP, Herndon DR, Ueti MW, Scoles GA, Kappmeyer LS, Suarez CE. Expression of 6-Cys Gene Superfamily Defines *Babesia bovis* Sexual Stage Development within *Rhipicephalus microplus*. PLoS One. 2016;11:e0163791.

47. Outchkourov NS, Roeffen W, Kaan A, Jansen J, Luty A, Schuiffel D, van Gemert GJ, van de Vegte-Bolmer M, Sauerwein RW, Stunnenberg HG. Correctly folded Pfs48/45 protein of Plasmodium falciparum elicits malaria transmission-blocking immunity in mice. PNAS. 2008;105:4301-4305.

48. Katzer F, McKellar S, Ferguson MA, d'Oliveira C, Shiels BR. A role for tertiary structure in the generation of antigenic diversity and molecular association of the *Tams1* polypeptide in *Theileria annulata*. Mol Biochem Parasitol. 2002;122:55-67.

49. Liu Y, Tewari R, Ning J, Blagborough AM, Garbom S, Pei J, Grishin NV, Steele RE, Sinden RE, Snell WJ. The conserved plant sterility gene HAP2 functions after attachment of fusogenic membranes in *Chlamydomonas* and *Plasmodium* gametes. Genes Dev. 2008;22:1051-1068.

50. Mori T, Hirai M, Kuroiwa T, Miyagishima S-y. The functional domain of GCS1-based gamete fusion resides in the amino terminus in plant and parasite species. PloS one. 2010;5:e15957.

51. Wong JL, Johnson MA. Is HAP2-GCS1 an ancestral gamete fusogen? Trends Cell Biol.2010;20:134-141.

52. Kangwanrangsan N, Tachibana M, Jenwithisuk R, Tsuboi T, Riengrojpitak S, Torii M, Ishino T. A member of the CPW-WPC protein family is expressed in and localized to the surface of developing ookinetes. Malar J. 2013;12:1.

53. Ghosh S, Ray DD, Vanlahmuaka, Das G, Singh NK, Sharma JK, Azhahianambi P. Progress in development of vaccine against *Hyalomma anatolicum anatolicum* -Indian scenario. Vaccine. 2008;26 Suppl 6:G40-G47.

54. Shakya M, Kumar B, Nagar G, de la Fuente J, Ghosh S. Subolesin: A candidate vaccine antigen for the control of cattle tick infestations in Indian situation. Vaccine. 2014;32:3488-3494.

Tables

Table 1

T. annulata ID	Chr	Product	Annotation	Signal peptide	GPI anchor	TMD	Macro to mero FC	Macro to piro FC	Mero to piro FC	T. parva ID	dNdS	Protein identity	Nucleotide identity
TA02580	3	Hypothetical protein	-			0	2.14	2.74	1.28	TP03_0040	0.0732	28.07	46.00
TA02585	3	Hypothetical protein	-			0	1.32	2.34	1.77	TP03_0039	0.0127	31.29	47.24
*TA03640**	3	Hypothetical protein	Sexual stage antigen (Pfam:PF07422)			0	3.55	4.14	1.17	TP03_0268	0.2529	64.17	73.18
TA03755	3	Sporozoite surface antigen (SPAG)	P67 sporozoite (Pfam:PF05642)			1	4.23	4.29	1.01	TP03_0287	0.3260	49.85	63.60
*TA10955**	4	Putative papain-family cysteine protease	Cysteine-type peptidase activity (GO:0008234)			0	3.42	4.61	1.35	TP04_0598	0.0904	85.50	83.48
TA13810	2	Putative ts-chitose type 23 kDa piroplasm surface-like protein	Orthologous to *T. sergenti* merozoite surface antigen			1	2.55	1.97	-1.29	TP02_0551	0.1638	83.41	84.57
TA13825	2	Hypothetical protein	-			1	12.40	11.00	-1.13	TP02_0553	0.4276	49.76	64.30
TA16005	2	Hypothetical protein	Domain of unknown function DUF529 (Pfam:PF04385)			1	1.87	3.09	1.65	TP02_0950	0.3317	62.59	75.61
TA16565	1	Hypothetical protein	-			0	1.60	2.27	1.41	TP01_1144	0.2704	74.68	79.69
*TA17050**	1	Merozoite-piroplasm surface antigen Tams1	Merozoite antigen (Pfam:PF02488)			1	2.62	2.64	1.01	TP01_1056	0.2751	72.86	77.38
TA17220	4	Hypothetical protein	Domain of unknown function DUF529 (Pfam:PF04385)			1	1.66	2.81	1.69	TP04_0030	0.3037	51.59	68.65
*TA20855**	1	Hypothetical protein	Similarity to *P. yoelii* (SWALL:EAA20932) and *P. falciparum* (SWALL:Q8IE86)			0	2.57	2.63	1.03	TP01_0412	0.085	80.43	77.72

Bioinformatic prediction of surface location of 12 candidates genes with significant fold change in gene expression levels between macroschizont and merozoite/piroplasm stages. Candidates selected for allelic sequencing are marked with an asterisk and protein and nucleotide identity are to the putative *T. parva* orthologue.

Figure Legends

Figure 1. Gene expression profile of 13 transmission-blocking candidates

Microarray expression profiles (A), (B) and (C) of *T. annulata* candidate genes in sporozoite, schizont, merozoite Day 4, Day 7, Day 9 and piroplasm. Expression is depicted on a $\log_2$ scale.

Figure 2. Protein alignments with related Apicomplexan genera

A. Alignment of the conserved papain family cysteine protease domain of Serine Repeat Antigen (SERA)-like Proteins from : *P. falciparum* (PF3D7_0207600), *P. reichenowi* (PRCDC_0206900), *T. orientalis* (TOT_040000333), *T. annulata* (*TA10995*), *T. parva* (TP04_0598).

B. Alignment of the highly conserved s48-45 superfamily 6-cysteine domain from sequences of *TA03640* homologues with conserved cysteine residues in green: *P. vivax* (PVP01_113600), *P. yoellii* (PYO3100), *P. chabaudi* (PCHAS_0111600), *T. annulata* (*TA03640*), *T. parva* (TP03_0268), *T. orientalis* (TOT_030000578), *B. bigemina* (BBBOND_0402900).

C. Alignment of the highly conserved 8-cysteine domain region of (revised) *TA20855* homologues with predicted signal peptides (blue), transmembrane helices (red) and cysteine residues (green) highlighted: *V. brassicaformis* (VBRA_17621), *T. gondii* (*TGME49_321580*), *P. falciparum* (*PF3D7_1322900*) *T. annulata*_revised (*TA20855*), *B. bovis* (*BBOV_IV006060*), *B. bigemina* (*BBBOND_0208520*).

Figure 3. qRT-PCR of candidate genes in tick stages

Quantitative RT-PCR expression analysis of RNA from *T. annulata* infected ticks generated at Day 2, Day 6, Day 10 and Day 15 post-detachment, relative to *T. annulata* merozoite Day 8 (calibrator) for: *Tams1* (*TA17050*); putative papain-family cysteine protease (*TA10955*); Pfs 48/45 6-cys domain encoding gene *TA03640* and 6-Cys like gene *TA20855*. * above (positive) or below (negative) error bars denote degrees of significant difference (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.0001$) between fold-change at a time-point relative to merozoite calibrator RNA.

Figure 4. Diversifying selection and B cell epitopes in candidate genes

dN/dS computed from allelic sequences of (a) *TA17050*; (b) *TA10955*; (c) *TA03460* and (d) *TA20855*, plotted against predicted B cell epitope regions (>1 amino acid, shaded areas) from GeneDB reference sequences for each protein. Epitopes with evidence of positive selection (peaks above 0) are shaded yellow, epitopes with no evidence of positive selection are shaded grey. For*TA20855* and *TA10955*, but not

TA17050 or *TA03640*, there was a statistically significant association between predicted epitopes and positive dN/dS, reflected in the figure by more yellow shaded areas, relative to grey.

Additional files

Additional file 1

Additional file 1.docx

Revised gene models for *T20855* and *TA19820* validated by RNA-seq reads, and alternate prediction of TM helices of *TA20855* using different software.

Additional file 2

Additional file 2.docx

qRT-PCR and allelic sequencing primers

Additional file 3

Additional file 3.docx

Transciptome data mined from EuPathDB for *Plasmodium* and *Toxoplasma* homologues of gene *TA20855*.

Additional file 4

Additional file 4.docx

***TA14250* encodes a second 6-cys (s48_45) domain protein, predicted to be expressed in the tick vector.**

Additional file 5

Additional file 5.docx

***TA09115* encodes the HAP2 domain found in proteins essential for gamete fusion, predicted to be expressed in the tick vector.**

Additional file 6

Additional file 6.docx

***TA19820* encodes a CPW-WPC domain protein, predicted to be expressed in the tick vector.**